Data Visualization: Assignment 3

Varnit Mittal
IMT2022025
Varnit.Mittal@iiitb.ac.in

Aditya Priyadarshi
IMT2022075
Aditya.Priyadarshi@iiitb.ac.in

Ananthakrishna K
IMT2022086
Ananthakrishna.K@iiitb.ac.in

I. Introduction

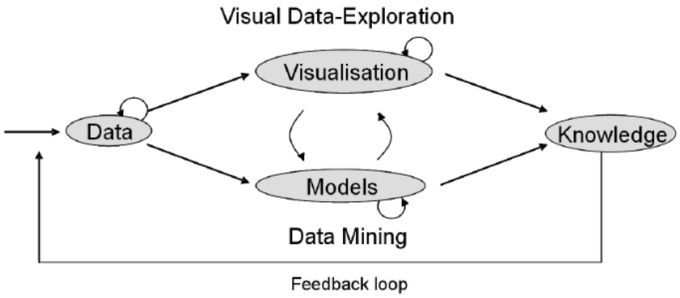

Fig. 1: Kiem et al. Visual Analytics Workflow, Image Courtesy: [2]

Figure 1 illustrates a typical visual analytics workflow, as discussed in [2]. In this study, we analyze the India census dataset, focusing on three major areas outlined below, in accordance with the framework proposed in [2].

1) **Task 1:** Analyzing Literacy-Driven Developmental Patterns
2) **Task 2:** Analyzing Literacy and Socio-Economic Dynamics in Female-Headed Households Through Census Data Visual Analytics
3) **Task 3:** Analyzing Workforce Demographics in India.

II. Task 1:Analyzing Literacy-Driven Developmental Patterns

The workflow illustrated in Figure 2 outlines a systematic approach to analyzing literacy-driven developmental patterns. Building on Task 3 from Assignment 1, it leverages initial visualizations to extract insights that guide the first iteration of the workflow. Based on these insights, areas for methodological refinement are identified and integrated into subsequent iterations. This iterative process follows the framework proposed by Kiem et al. [2], ensuring continuous improvement through a feedback-driven approach.

A. Additional Dataset

For this assignment, the dataset sourced from the official Indian Government's Census dataset [3] was utilized, as it offers a more comprehensive and reliable set of data for analysis. The dataset previously used in Assignment 1 [1] had several inconsistencies, particularly where the sum of literates, illiterates, and individuals who did not disclose information did not match the total population, undermining the dataset's reliability. Additionally, the initial dataset lacked important data on urbanization, which was crucial for the analysis in this task. These issues led to the adoption of the more accurate and complete dataset from [3], which includes critical columns like urbanization data and ensures a higher level of consistency for the analysis. For more insightful comparisons, data from both 2001 and 2011 were used, enabling a more robust analysis of changes over the decade, particularly with regard to literacy and urbanization trends.

B. Preprocessing

The additional dataset provided data in multiple tables, each corresponding to specific components of the Census. To gain deeper insights, these tables were merged. We used Microsoft Excel and Python's Pandas library for efficient data manipulation and analysis. The keys steps in the preprocessing include –

1) **Aggregation**: The dataset was initially structured at the district level. For state-wise analysis, the data was aggregated at the state level by summing district values for each state.
2) **Feature Engineering**: New columns were derived to better capture key indicators. The following transformations were performed:
   a) Literacy Rate: Proportion of literates to total population.
   b) Higher Education Rate: Proportion of the population with higher or above education.
   c) Higher Education Rate (Male/Female): Gender-specific rates of higher or above education.
   d) Literacy Rate (Male/Female): Gender-specific literacy rates.
   e) Urbanization Index: Ratio of urban population to total population.
   f) Population Growth Rate: Growth in total population over time.
   g) Population Density Growth Rate: Change in population density over time.

   - **Note:** Higher and above education includes people with technical or non-technical diplomas, secondary education, higher education, and graduate or above education.
3) **Data Cleaning:** To ensure data integrity, the following checks were conducted:

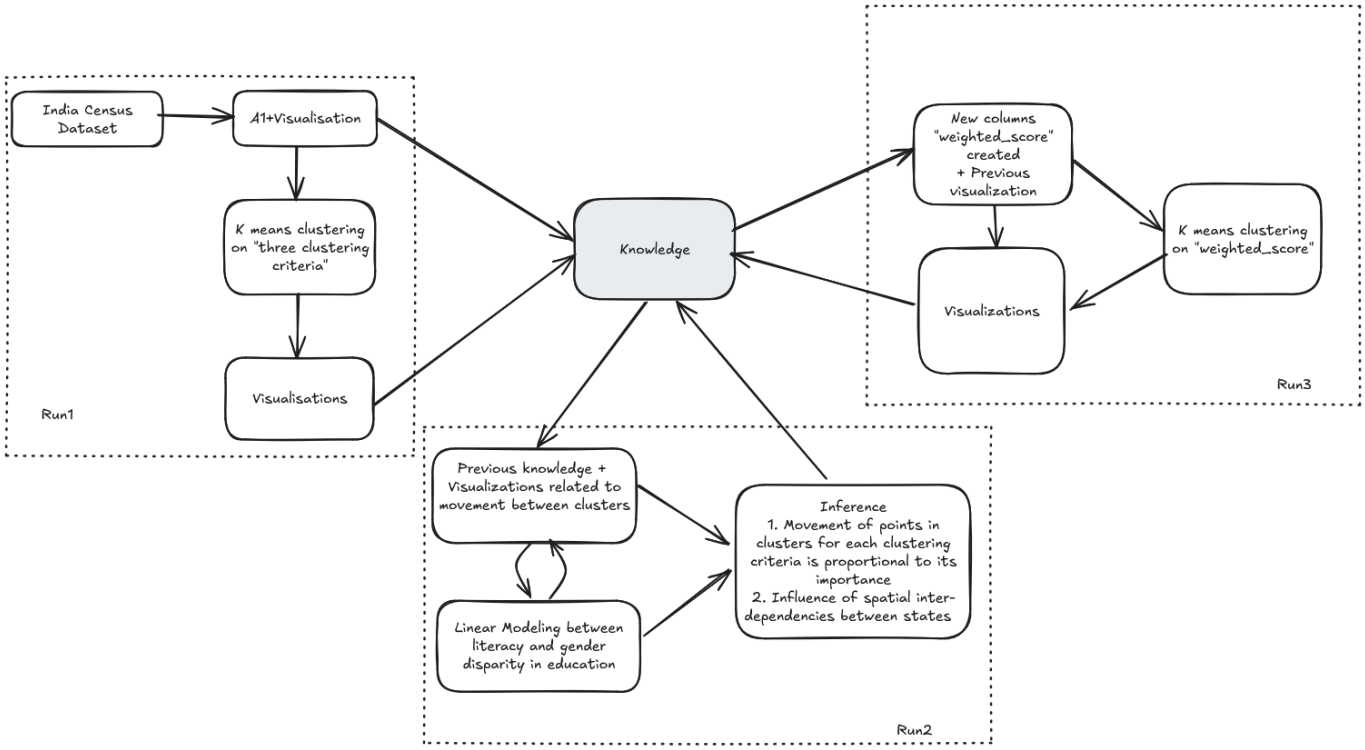

Fig. 2: Visual analytics workflow utilized in Task 1, illustrating iterative processes as described in Kiem et al [2]. Each rectangle with dotted side represents a complete workflow cycle.

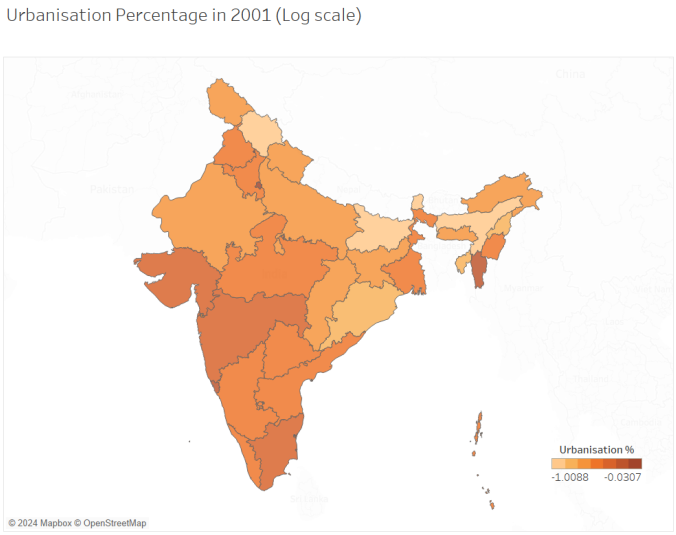

Fig. 3: Urbanization Percentage in 2001 (Log Scale)

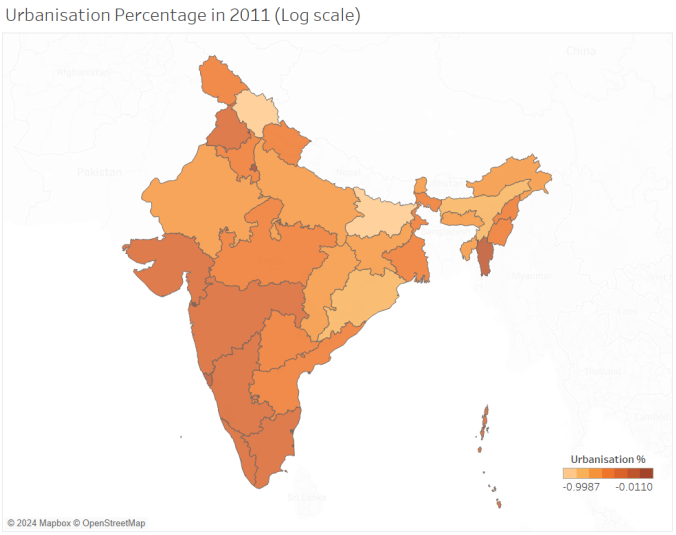

Fig. 4: Urbanization Percentage in 2011 (Log Scale)

- **Data Type Validation:** Verified that all columns (obviously except State Name) had appropriate numeric data types.
- **Handling Missing Values:** Checked for missing (NaN) values, but none were found.

C. Visualizations and inferences from Assignment 1

In this section, we present the visualizations related to urbanization and literacy, which provide an overview of the patterns across different years (2001 and 2011). These visualizations offer valuable insights into regional disparities in urbanization and educational attainment.

1) **Urbanization Percentage in 2001 and 2011:** Figures 3

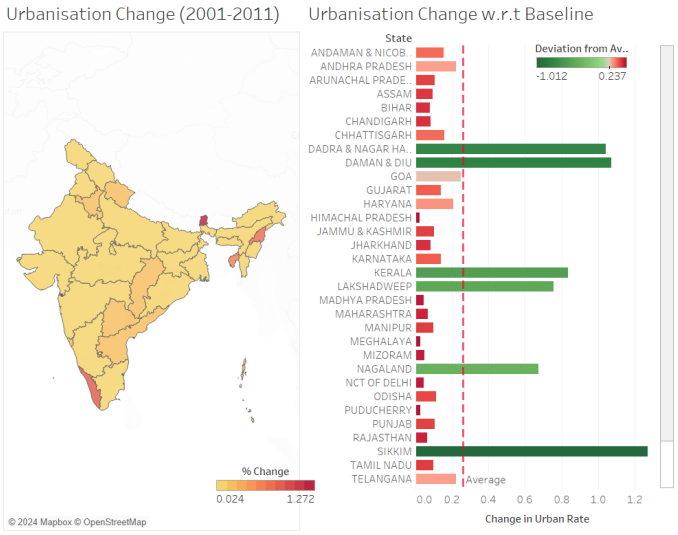

Fig. 5: Urbanization Change from 2001 to 2011 (Relative to Baseline)

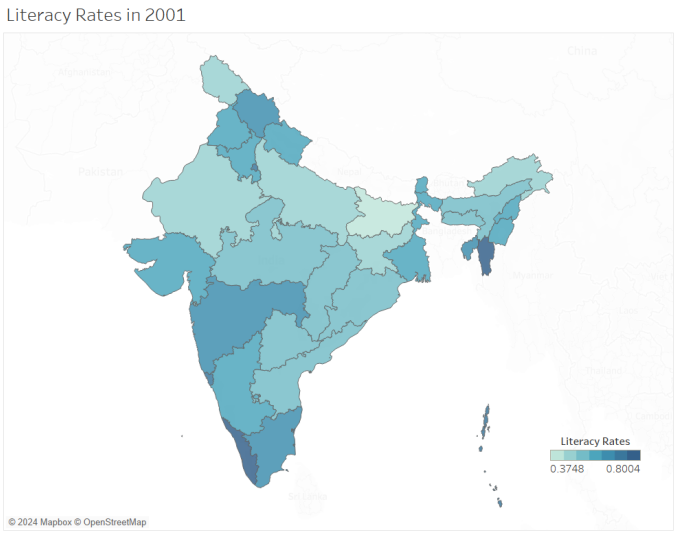

Fig. 6: Literacy Rates in 2001

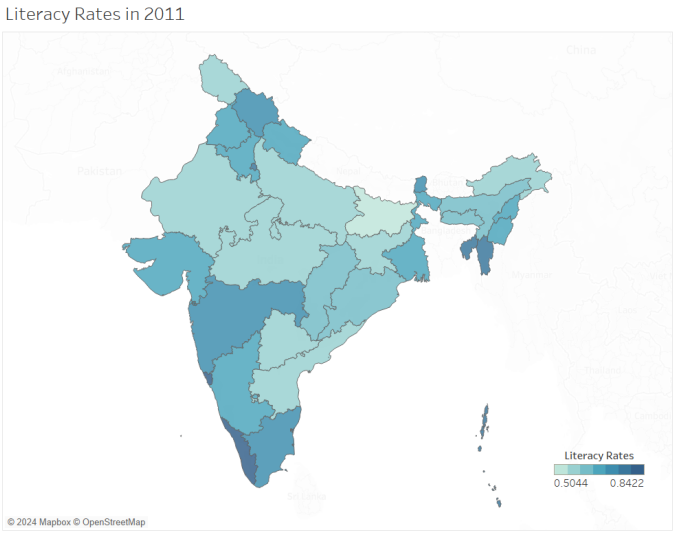

Fig. 7: Literacy Rates in 2011

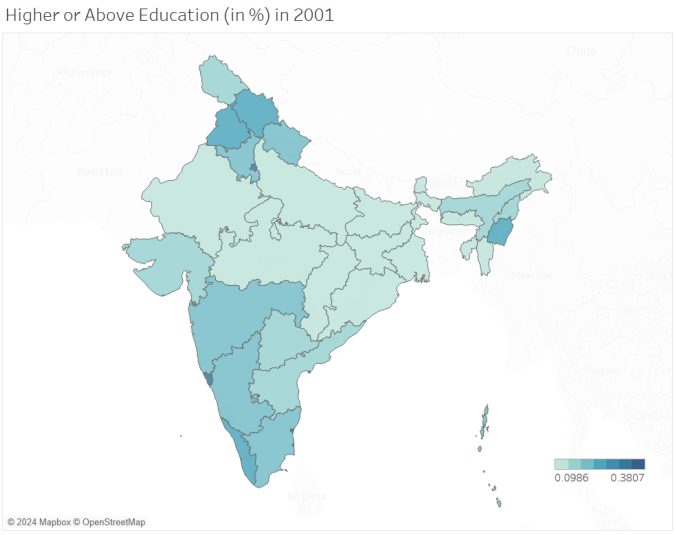

Fig. 8: Percentage of Population with Higher Education and Above in 2001

and 4 display urbanization percentages using choropleth maps with a log scale to enhance clarity. The use of a log scale was applied in these visualizations to compress the range of values, which allows for a clearer comparison across states, particularly in areas with extremely low or high urbanization percentages. Southern states such as Maharashtra, Tamil Nadu, and Gujarat have higher urbanization rates, while states like Bihar, Uttar Pradesh, and Assam exhibit lower urbanization.
Inference: A clear regional divide is seen, with southern states more urbanized, while northern and northeastern states remain largely rural.

2) **Urbanization Change (2001 to 2011):** Figure 5 shows the urbanization change, highlighting states like Sikkim, Daman and Diu, Nagaland, and Dadra and Nagar Haveli with substantial urban growth, compared to slower-growing states like Himachal Pradesh, Rajasthan, and Madhya Pradesh.
Inference: States with rapid urbanization growth indicate improved infrastructure or migration trends, while slower growth may reflect limited urban development.

3) **Literacy and Higher Education Rates:** Figures 6, 7, 8, and 9 illustrate literacy and higher education rates using choropleth maps. Southern states such as Maharashtra, Tamil Nadu, and Gujarat show higher rates, while states like Bihar, Uttar Pradesh, and Assam exhibit lower rates.
Inference: Regional disparities in literacy and education levels are evident, with southern states outperforming

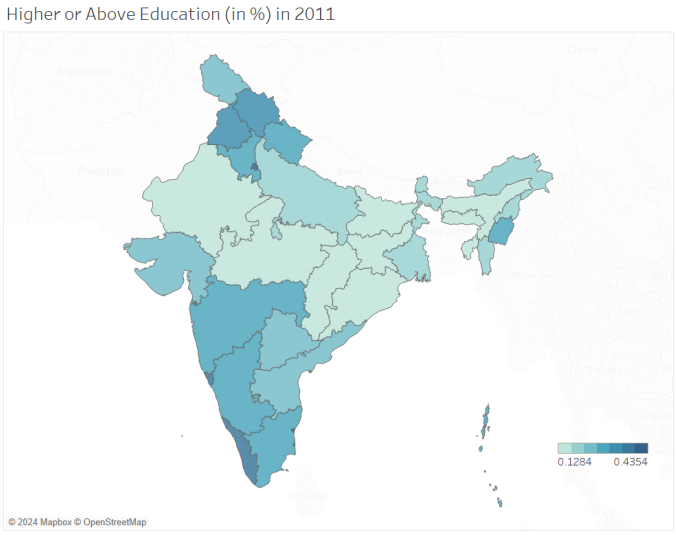

Fig. 9: Percentage of Population with Higher Education and Above in 2011

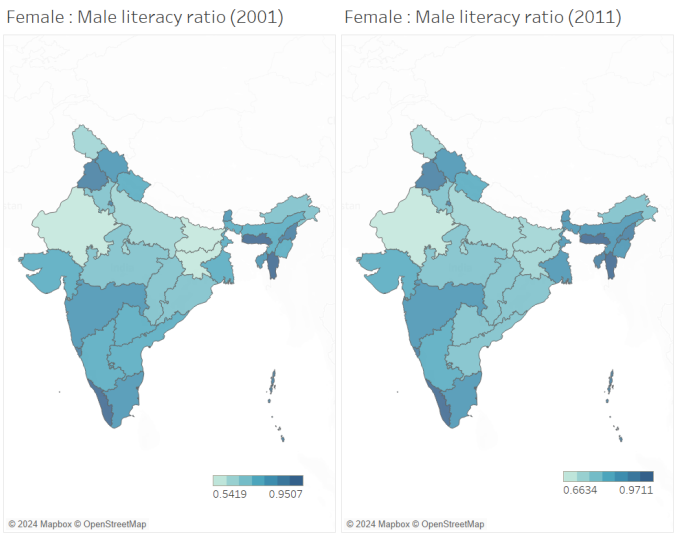

Fig. 10: Female to Male Literacy Ratios

others, potentially due to better educational policies and resources.

4) **Female-to-Male Literacy Ratio:** Figure 10 illustrates the female-to-male literacy ratio using choropleth maps. Southern states like Kerala, Tamil Nadu, and Maharashtra exhibit darker regions, indicating greater gender parity in literacy. Conversely, states like Bihar, Uttar Pradesh, and Rajasthan show lighter regions, reflecting significant disparities in literacy rates between females and males.
Inference: Regional disparities in literacy and gender parity are evident, with southern states outperforming others, likely due to more inclusive educational policies and better access to resources for women.

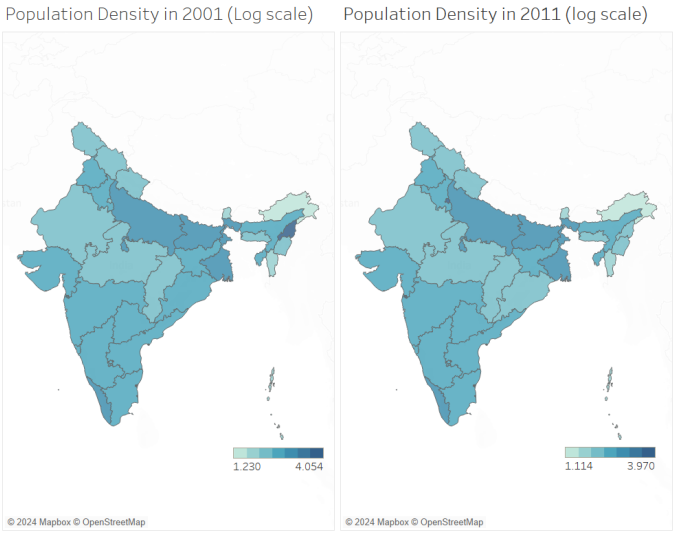

Fig. 11: Population Density

5) **Population Density:** Figure 11 represent population density using choropleth maps. States such as Bihar, West Bengal, and Uttar Pradesh exhibit darker regions, indicating higher population density, while states like Arunachal Pradesh, Sikkim, and Himachal Pradesh show lighter regions, reflecting lower population density.
Inference: The disparity in population density highlights significant regional differences, with densely populated states facing greater challenges in resource allocation and infrastructure development compared to sparsely populated regions.

D. The Visual Analytics Workflow - First Run

1) Data: The dataset utilized for the first run of the Visual Analytics Workflow consisted of several columns, including `Area`, `Total/Urban/Rural`, `Population`, `Illiterate`, `Literate`, `No Formal Education`, `Below Primary`, `Primary`, `Middle`, `Secondary`, `Higher`, `Graduate`, `Area of State`, and `Population Density` for males, females, and total population. The exact column names are provided in [4].

Since urban and rural population data was essential for the analysis, the urban row for each state was aggregated into a new column labeled `UrbanPop`. Subsequently, the individual rows for urban and rural populations were removed, which was carried out using Microsoft Excel. In addition to this data aggregation, all preprocessing steps, including data cleaning and feature engineering, were also performed to ensure the dataset's integrity and readiness for analysis.

2) Data Mining (using Machine Learning model): **Clustering Model:**

For the first run, a K-Means clustering model was employed [5]. This method was chosen for its simplicity, scalability, and effectiveness in partitioning data into distinct groups based on similarity.

Clustering was applied to the following three sets of features for both years, 2001 and 2011:

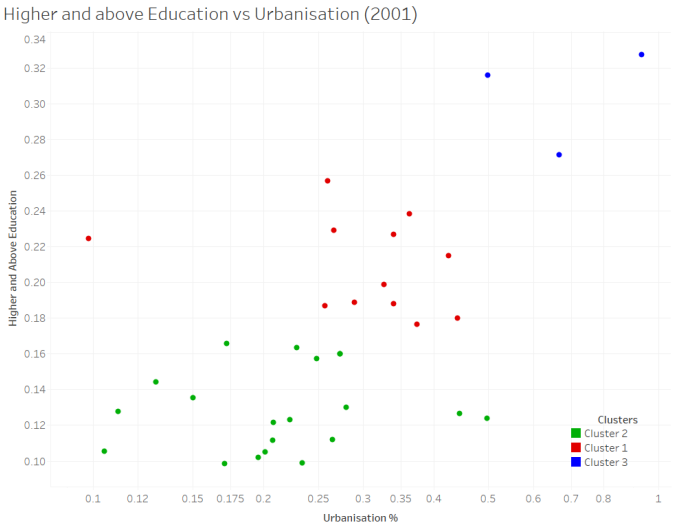

Fig. 12: Clustering of Higher and Above Education vs. Urbanisation (2001)

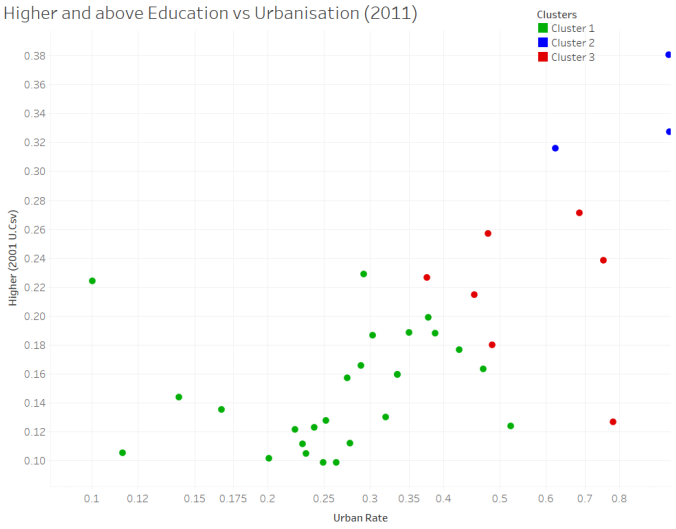

Fig. 13: Clustering of Higher and Above Education vs. Urbanisation (2011)

1) **Literacy Rate and Population Density:** Clustering was performed using the literacy rate and population density. The goal was to identify patterns of urbanization and education distribution based on these two factors.
2) **Higher Education and Urbanization:** In this case, clustering was based on the higher education rate and urbanization level for each state. This combination reflects the relationship between education and urban development.
3) **Overall Literacy vs. Female-to-Male Literacy:** This clustering was done using the overall literacy rate and the ratio of female to male literacy to capture gender-related disparities in literacy.

In each case, the optimal number of clusters (k) was chosen to be 3, based on visual inspection (as suggest by Andrew Ng [6]) and domain knowledge, which indicated that three clusters provided meaningful distinctions in each of the visualizations. The choice of clustering features was further validated by noticing that the darker regions in the initial visualizations (such as those depicting urbanization or literacy) corresponded with similar dark regions in other visualizations, suggesting coherence across different data aspects.

For generating the clusters, Tableau was used, which internally employs the K-Means clustering algorithm with a variance-based partitioning method.

3) Visualizations: In this step, scatter plots were created for each of the three clusters, with colors indicating each cluster. These visualizations were chosen because they effectively highlight the distribution of data points across different clusters, making it easier to visually interpret the relationships between the clustered features. Scatter plots allow for a direct comparison of the clustering results for key variables. To gain additional insights and context, the visualizations outlined in the section II-C were also utilized, providing a deeper understanding of the clustering patterns and their domain relevance.

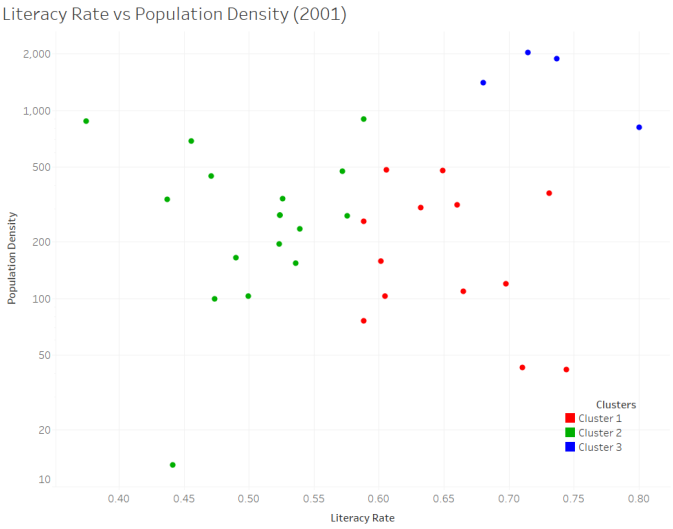

Fig. 14: Clustering of Literacy Rate vs. Population Density (2001)

4) Knowledge:

1) **Higher Education and Urbanization:** Upon comparing the choropleth maps of higher education rates in Figures 8 and 9 with those of urbanization rates in Figures 3 and 4 and its changes (Figure 5), we observed a significant overlap in the darker regions, indicating a strong correlation between higher education rates and urbanization levels. When clustering the states based on higher education and urbanization (as shown in Figures

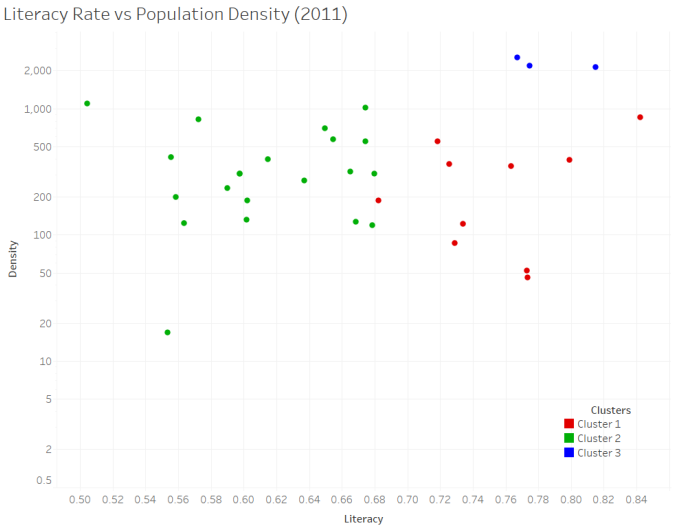

Fig. 15: Clustering of Literacy Rate vs. Population Density (2011)

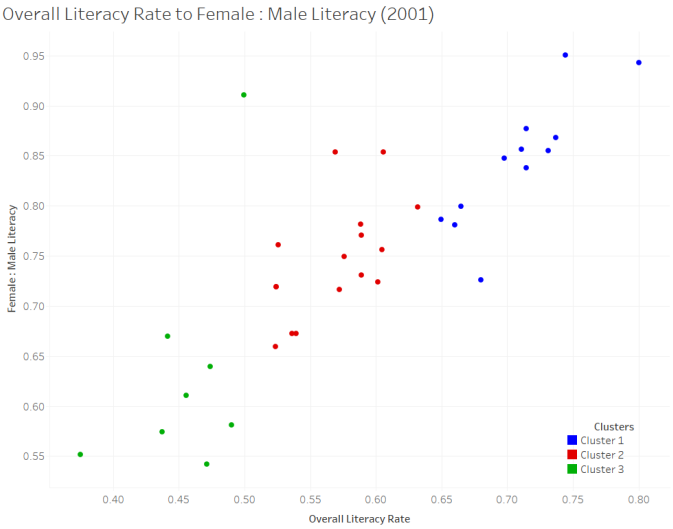

Fig. 16: Clustering of Clustering of Overall Literacy Rate vs. Female-to-Male Literacy Rate (2001)

13 and 13), the resulting clusters revealed the following insights:

- **2001, Cluster 1:** Predominantly states with poor outcomes in both higher education and urbanization.
- **2011, Cluster 1:** States that showed substantial improvements by 2011, reflecting better outcomes in both higher education and urbanization.
- **2011, Cluster 2:** States from 2001 that did not exhibit the same level of improvement by 2011, still lagging in terms of education and urbanization.
- **2001, Cluster 3:** States that performed well in 2001 in terms of both higher education and urbanization.
- **2001, Cluster 2 $\approx$ 2011, Cluster 3:** Union Territories (UTs) that demonstrated positive outcomes, though they might be outliers in the overall pattern.

2) **Literacy Rate and Population Density:** A comparison of the choropleth maps for literacy rates in Figures 6 and 7 with the population density maps revealed that regions with higher literacy rates often coincide with areas of lower to moderate population density (in Figure 11). By clustering the states based on literacy rates and population density for both 2001 and 2011 (as shown in Figures 14 and 15), the following patterns were identified:

- **2001 / 2011, Cluster 1:** These states exhibited better performance in terms of literacy rates and population density, reflecting higher levels of development and education infrastructure.
- **2001 / 2011, Cluster 2:** States that performed worse in terms of literacy rates and population density, often corresponding to regions with underdeveloped education systems and low urbanization.
- **2001 / 2011, Cluster 3:** Outlier states or Union Territories (UTs) that displayed exceptionally high

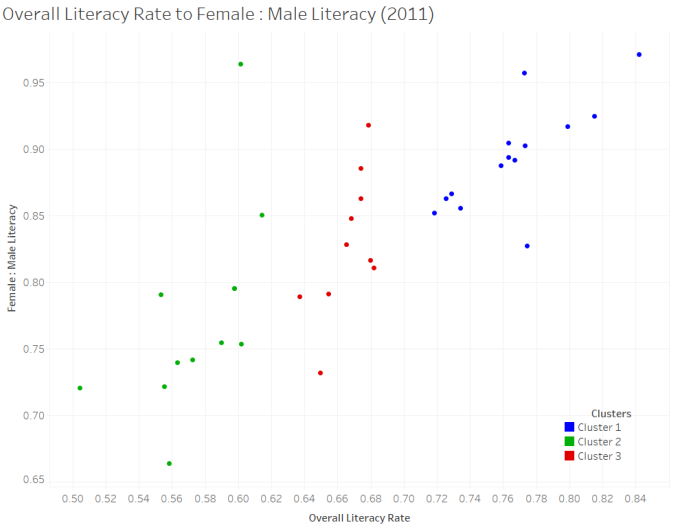

Fig. 17: Clustering of Clustering of Overall Literacy Rate vs. Female-to-Male Literacy Rate (2011)

literacy rates relative to their population density, indicating over-performance when compared to national trends.

3) **Overall Literacy vs. Female-to-Male Literacy:** The comparison of choropleth maps for overall literacy rates (Figures 6 and 7) with the clustering results based on female-to-male literacy ratios (Figures 10 highlights the correlation between gender parity in literacy and overall literacy levels. The clustering of states based on these metrics for 2001 and 2011 (Figures 16 and 17) revealed the following insights:

- **Cluster 1:** States with the best overall performance, characterized by high literacy rates and near-parity

between female and male literacy. These states reflect significant strides in both educational access and gender equality.

- **Cluster 2:** States with moderate performance, showing reasonable overall literacy rates but lagging slightly in achieving gender parity in literacy.
- **Cluster 3:** States with the lowest performance, marked by low overall literacy rates and substantial gaps in female-to-male literacy ratios, indicating entrenched gender disparities in educational access.

The hierarchy of the clusters follows the order of performance: **Cluster 1** $>$ **Cluster 2** $>$ **Cluster 3**, with states in Cluster 1 showcasing a balanced and advanced education system relative to the others.

5) Feedback: The clustering results reveal three distinct categories of clusters for each analysis: good performers, average performers, and poor performers. Additionally, the clusters are not static over time; states in a particular cluster in 2001 do not necessarily remain in the same cluster in 2011. This temporal variability indicates significant shifts in socio-economic dynamics.

To better understand **State Transitions** and **Cluster Dynamics**, the cluster assignments from the analyses were added as new columns to the dataset. This enables the identification of movers—states that transitioned to higher-performing clusters and stagnators—states that remained in or regressed to lower-performing clusters. Through this, patterns of socio-economic development across regions can be effectively spotted and analyzed, providing valuable insights into regional disparities and developmental trajectories.

E. The Visual Analytics Workflow - Second Run

1) Data: In the second run of the Visual Analytics Feedback Loop, the three cluster columns from the clustering step in Tableau were added to the dataset. These columns represent the cluster assignments for each state based on the three analyses: literacy and population density, higher education and urbanization, and overall literacy versus female-to-male literacy. The updated datasets were then exported as CSV files, named `2001_iter1.csv` and `2011_iter1.csv`, for further analysis.

2) Data Mining (using Machine Learning and Statistics): To explore the relationship between the female-to-male literacy ratio and overall literacy rate, a **linear regression model** was constructed. This approach helps assess whether a significant linear dependency exists between these two variables.

- **Linear Regression Setup:**
  - **Predictor Variable (X):** Female-to-male literacy ratio ($\frac{\text{literacy\_female}}{\text{literacy\_male}}$).
  - **Target Variable (Y):** Overall literacy rate (literacy).

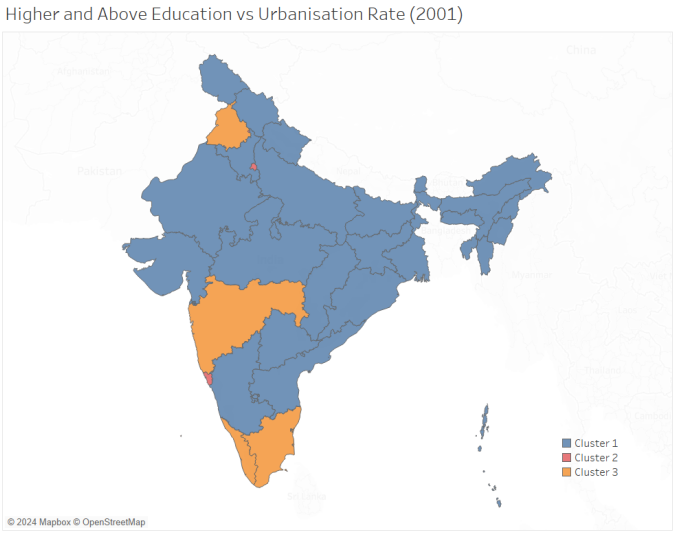

Fig. 18: Higher Education and Urbanization Rates Boundaries (2001)

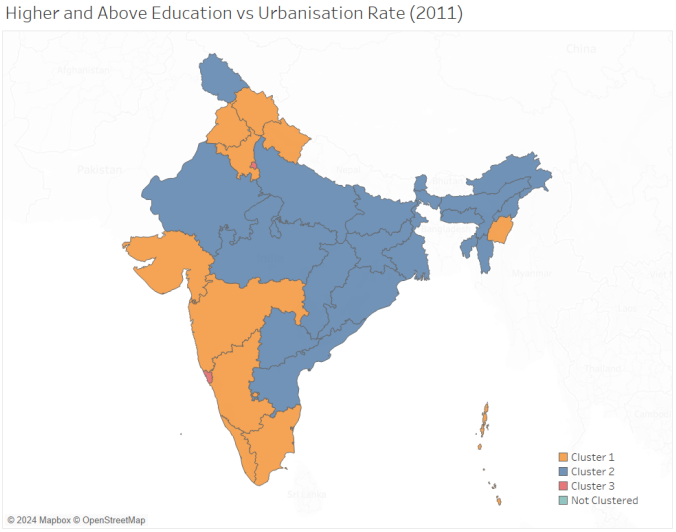

Fig. 19: Higher Education and Urbanization Rates Boundaries (2011)

  - The dataset was split into training and testing sets (80%-20%) for training the model and evaluating its predictive performance.
- **Key Statistical Measures:**
  - **Coefficient** (β) and **Intercept** (α)**:** The regression model estimates these values to fit a line to the data.
  - **Mean Squared Error (MSE):** This metric quantifies the average squared differences between predicted and actual values in the test set, assessing model accuracy.
- **T-Test for Coefficient Significance:**
  - **Null Hypothesis** (H_0)**:** The predictor variable

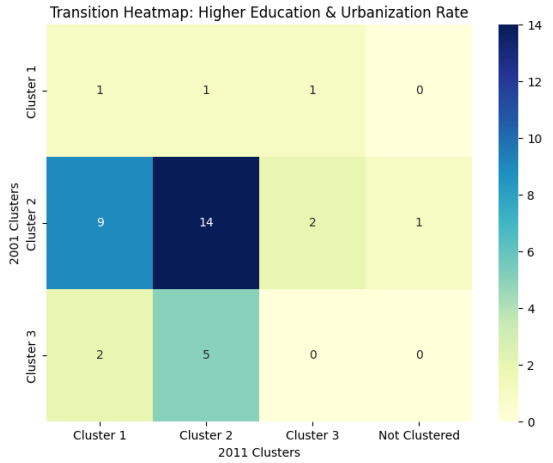

Fig. 20: Cluster Transition Heatmap (2001-2011): Comparison of Higher Education and Urbanization Rates

(X) has no significant linear relationship with the target variable (Y), i.e., $\beta = 0$.

- **Alternative Hypothesis (H_1):** A significant linear relationship exists, i.e., $\beta \neq 0$.
- The test uses the following:
  - **Standard Error of the Coefficient (SE_β):** Measures the variability of the coefficient estimate.
  - **T-Statistic:** Calculated as $t = \frac{\beta}{SE_\beta}$, indicating how many standard errors the coefficient is away from zero.
  - **P-Value:** Derived from the t-distribution, indicating the probability of observing a t-statistic as extreme as the one computed, under H_0.

- **Interpretation:**
  - If p-value < 0.05, the null hypothesis is rejected, indicating a statistically significant linear relationship.
  - Otherwise, no significant relationship is concluded.

Results: For both 2001 and 2011, the regression analysis revealed a significant linear relationship between the female-to-male literacy ratio and overall literacy rate.

- **2001 Results:**
  - **Coefficient:** 0.8175
  - **T-Statistic:** 9.426
  - **T-Test P-Value:** 7.15×10^{-10}
  - A significant linear relationship exists between the female-to-male literacy ratio and overall literacy in 2001.
- **2011 Results:**
  - **Coefficient:** 0.9687
  - **T-Statistic:** 8.088
  - **T-Test P-Value:** 1.44×10^{-8}
  - A significant linear relationship exists between the female-to-male literacy ratio and overall literacy in 2011.

To evaluate the coherence of the clustering results, **pairwise correlations** were computed for the percentage changes in critical socio-economic metrics between 2001 and 2011: **Correlated Metrics:**

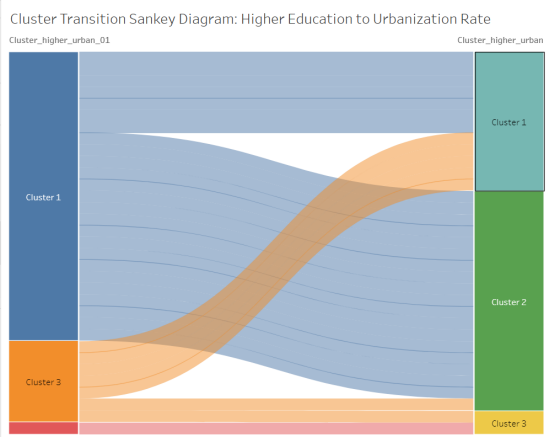

Fig. 21: Cluster Transition Sankey Diagram (2001 (Left) - 2011 (Right)) Higher Education and Urbanization Rates

- **Urbanization Rate vs. Higher and Above Education Rate:** Correlation = 0.623 (Moderate positive correlation, indicating that areas with higher urbanization tend to have improved higher education rates).
- **Population Density vs. Literacy Rate:** Correlation = -0.384 (Weak negative correlation, suggesting population density and literacy are inversely related).
- **Female-to-Male Literacy Ratio vs. Overall Literacy Rate:** Correlation = 0.773 (Strong positive correlation, showing that areas with better gender parity in literacy also have higher overall literacy).

3) Visualizations: In the second iteration of the visual analytics feedback loop, a series of visualizations were created to further explore the relationships between socio-economic indicators, cluster dynamics, and transitions. These visualizations provide valuable insights into the performance and movement of states across clusters over time. The visualizations presented are as follows:

a) **Linear Regression Fit for Overall Literacy vs. Female-to-Male Literacy Ratio (2001, 2011):** Figure 22 displays scatter plots of the relationship between the overall literacy rate and the female-to-male literacy ratio for the years 2001 and 2011. A linear regression line has been fitted to the data points, highlighting the linear relationship between the two variables. These visualizations clearly show how the female-to-male literacy ratio correlates with overall literacy across the different

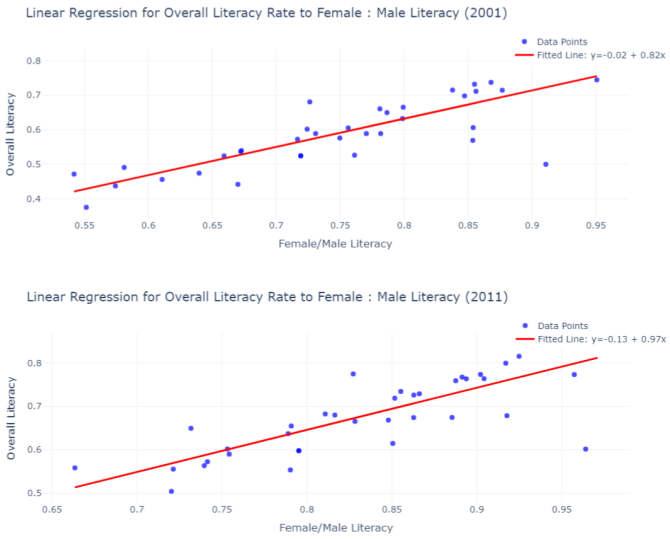

Fig. 22: Linear Regression fit for Overall to Female to Male Literacy

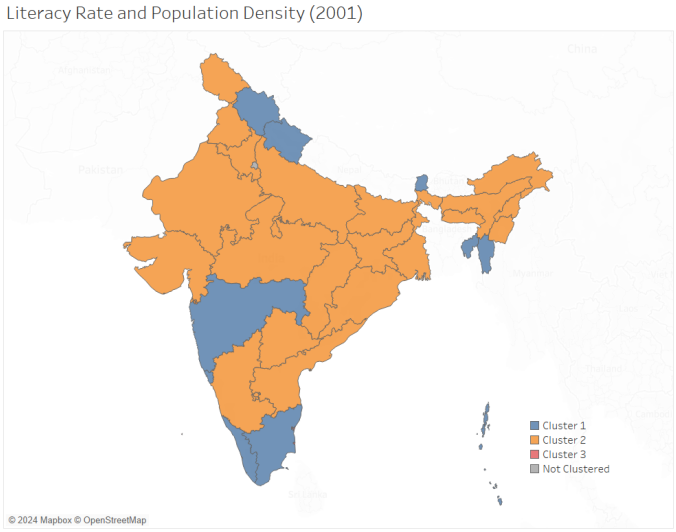

Fig. 23: Literacy Rates and Population Density Boundaries (2001)

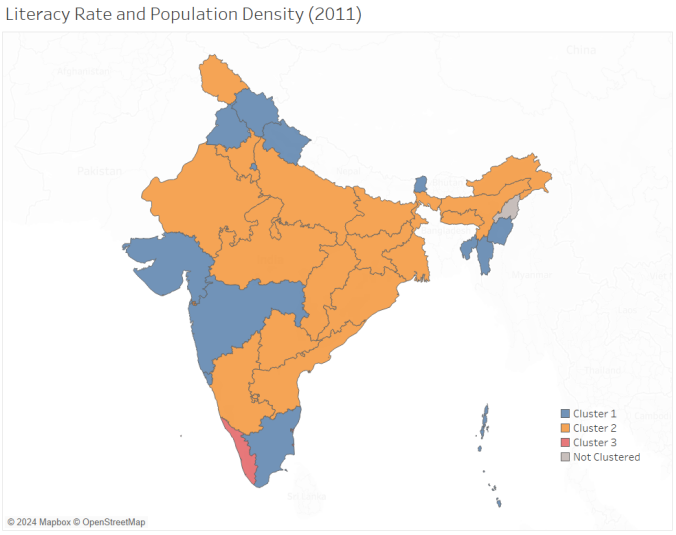

Fig. 24: Literacy Rates and Population Density Boundaries (2011)

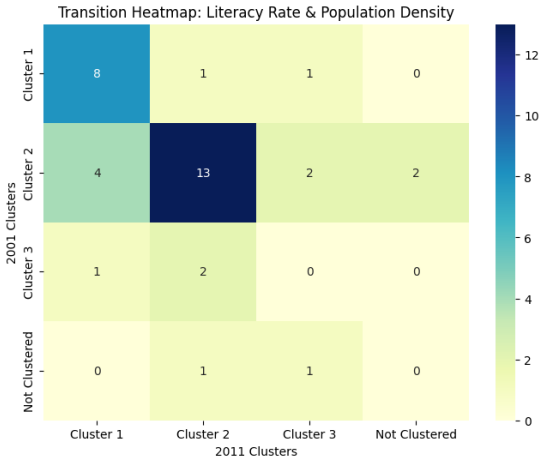

Fig. 25: Cluster Transition Heatmap (2001-2011): Comparison of Literacy Rates and Population Density

states. The line fit further reinforces the positive correlation between the two variables, which was statistically confirmed through regression analysis.

b) **Cluster Transition Heatmap (2001-2011):** Figures 20, 25 and 29 present heatmaps illustrating the transition of states between clusters in 2001 and 2011 (for all three clustering criteria discussed above in II-D4). The heatmap shows the three clusters on both the x-axis (representing 2011) and the y-axis (representing 2001). Each cell in the heatmap indicates the number of states that transitioned into a particular cluster from 2001 to 2011. This visualization provides a clear view of the shifts in state performance over the decade, revealing how states moved between the "good," "average," and "bad" performance clusters. By examining these transitions, we gain a deeper understanding of the regional dynamics and socio-economic development across India.

c) **Sankey Diagram of Cluster Movements:** Figures 21, 26 and 30 show Sankey diagrams that depict the movement of states between clusters for all three clustering criteria as discussed in II-D4. These diagrams visually represent the flow of states from one cluster to another, with the width of the flow proportional to the number of states transitioning between clusters. The Sankey diagrams facilitate a more intuitive understanding

of how clusters evolve and highlight any significant shifts in regional development patterns over time.

d) **Cluster Boundaries for 2001 and 2011:** Figures 18, 19, 24, 23, 27 and 28 illustrate the boundaries of the clusters for the years 2001 and 2011 for all three clustering criteria (II-D4). These maps show the geographic distribution of states within each cluster, offering a visual representation of the spatial distribution of performance across regions. By comparing the boundaries of clusters in 2001 and 2011, it is possible to assess whether regional disparities in socio-economic development have been addressed over time, as well as identify areas that have seen significant improvements or stagnation.

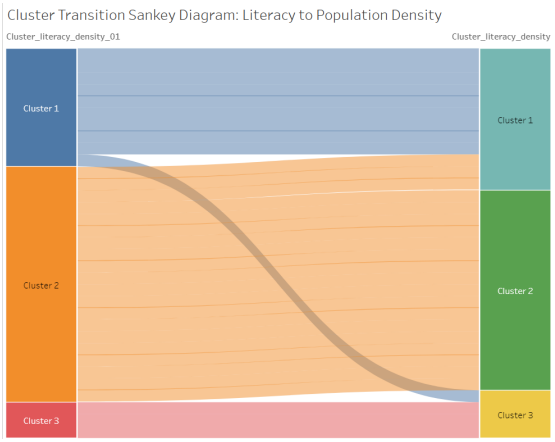

Fig. 26: Cluster Transition Sankey Diagram (2001 (Left) - 2011 (Right)) Literacy Rates and Population Density

Each of these visualizations serves as an essential tool for interpreting the results of the clustering and regression analyses. By combining these visuals, we are able to assess the impact of various socio-economic factors on literacy and urbanization, explore how states transitioned across clusters, and evaluate the effectiveness of regional development policies. The visualizations collectively offer a comprehensive view of the socio-economic landscape in India and highlight key areas for further investigation and policy intervention.

4) Knowledge:

a) From the insights provided in II-D4, the dynamics of cluster transitions for higher education and urbanization reveal distinct patterns of socio-economic development across states:

- In 2001, **Cluster 1** primarily represented states with poor performance in both higher education and urbanization. On 2011 map, cluster 1 includes states that demonstrated substantial improvements in these metrics, indicating that some states successfully progressed in higher education and its urbanization impact.

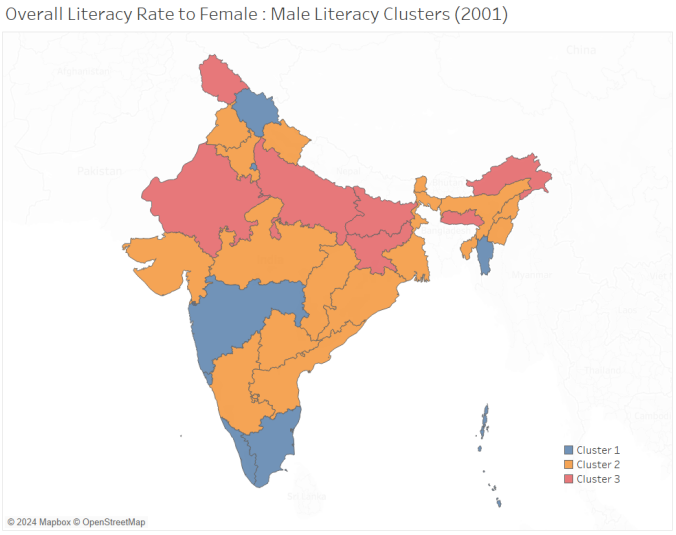

Fig. 27: Female : Male Literacy and Overall Literacy Rates Boundaries (2001)

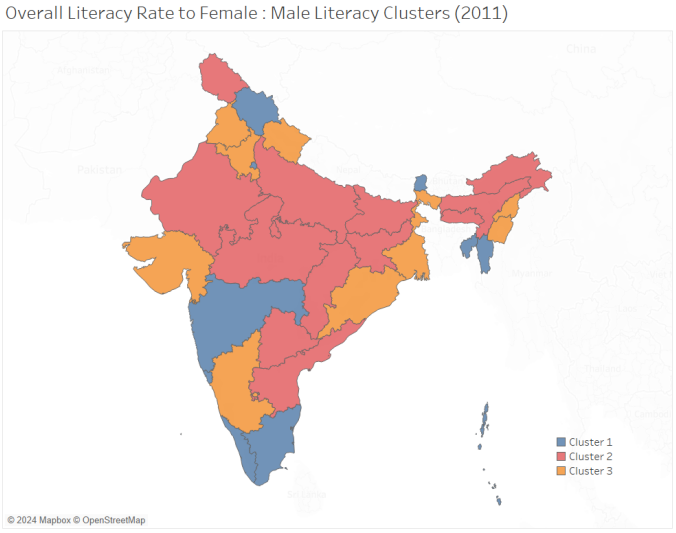

Fig. 28: Female : Male Literacy and Overall Literacy Rates Boundaries (2011)

- **Cluster 2** in 2011 predominantly consisted of states that were previously part of **Cluster 1** in 2001 but failed to achieve significant improvement over the decade. This cluster represents regions that remained relatively stagnant in terms of socio-economic progress.
- **Cluster 3** in 2001 included states that performed well in both metrics, such as Tamil Nadu, Maharashtra, Kerala, and Punjab. By 2011, these states continued to excel, influencing neighboring states such as Karnataka, Himachal Pradesh, Gujarat, Uttarakhand, and Haryana, which transitioned into **Cluster 1** alongside the original

Cluster 3 states. This indicates that the positive impact of higher education and urbanization in these high-performing states acted as a regional center of influence, driving development in surrounding areas.

- A notable observation is that Union Territories (UTs) and smaller states, such as Puducherry and Lakshadweep, represented by **Cluster 3** in 2001, did not propagate the benefits of higher education and urbanization to neighboring regions. These regions remained geographically and socio-economically isolated, leading to a minimal ripple effect in their surroundings.
- In contrast, a large portion of eastern, northeastern, and central states, such as Uttar Pradesh, West Bengal, Bihar, Jharkhand, and other northeastern states, remained in **Cluster 1** across both years. This suggests that these regions were unaffected by the urbanization-led development observed elsewhere, highlighting a persistent lack of progress in higher education and urbanization.

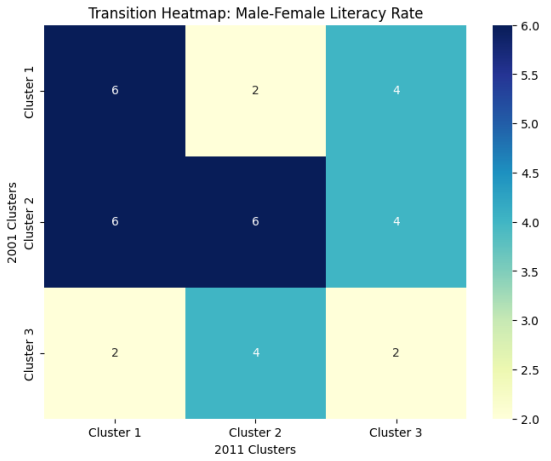

Fig. 29: Cluster Transition Heatmap (2001-2011): Comparison of Female : Male Literacy and Overall Literacy Rates

A critical insight from these transitions is the role of **spatial neighborhood effects** in influencing cluster movements. States within the same clusters tend to exhibit geographic adjacency rather than being dispersed across the country. This suggests that regional proximity and socio-economic linkages play a significant role in propagating developmental impacts. States such as Tamil Nadu, Maharashtra, and Kerala acted as anchors, spreading the positive effects of higher education and urbanization to neighboring regions. Conversely, regions without strong socio-economic neighbors, such as northeastern and central India, largely remained stagnant, unaffected by these trends.

This observation underscores the importance of considering spatial influences and inter-state collaborations when designing policies to foster regional development. While state-level policies are essential, creating interlinked frameworks that leverage neighborhood effects could accelerate progress in under-performing regions.

b) The transitions in clusters based on literacy rates and population density highlight regional disparities and limited mobility:

- **Cluster 1**, representing high-performing states in both 2001 and 2011, largely remained stable. Additional states, such as Gujarat, Punjab, and Manipur, transitioned into this cluster by 2011, likely influenced by their proximity to high-performing neighbors like Maharashtra, Haryana, Mizoram, and Tripura.
- **Cluster 2**, comprising states with average performance, saw movement, with the majority of states remaining stagnant, indicating a lack of significant improvement over the decade.
- **Cluster 3**, representing outliers or over-performers (mostly composed of UTs), remained relatively small and unchanged, signifying the persistence of exceptional literacy and density patterns in these states.

  These patterns suggest that progress in literacy and population density tends to follow regional proximity, with high-performing states influencing their immediate neighbors. However, the limited transitions across clusters underscore the entrenched disparities and the need for targeted interventions to uplift underperforming regions.

c) A similar pattern of cluster transitions is observed for the relationship between the female-to-male literacy ratio and overall literacy rates:

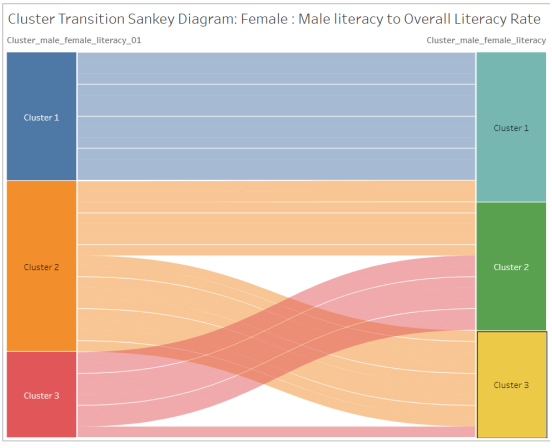

Fig. 30: Cluster Transition Sankey Diagram (2001 (Left) - 2011 (Right)) Female : Male Literacy and Overall Literacy Rates

- **Cluster 1**, representing states with the highest levels of gender parity and overall literacy, maintained consistency from 2001 to 2011. States such as Kerala, Mizoram, and Goa continued to dominate this cluster, with minimal additions, emphasizing their leadership in equitable literacy development.
- **Cluster 2**, comprising moderately performing states, showed limited transitions, with most states maintaining their positions, reflecting slow progress in reducing gender disparities in literacy.
- **Cluster 3**, denoting low-performing states, experienced little movement, highlighting entrenched gender gaps and low overall literacy rates in these regions.

  As with literacy rates and population density, the movement across clusters suggests that states with high gender parity in literacy influence neighboring regions positively, but significant barriers persist in uplifting under-performing areas. This underscores the need for targeted, region-specific policies to address both gender inequities and overall literacy deficits.

d) The heatmaps presented in Figures 20, 25 and 29 reveal varying levels of state transitions across clustering criteria. Certain criteria, such as those based on literacy rates and population density, exhibit relatively stable cluster memberships over time, while others, like higher education and urbanization rates, show significant transitions between clusters. This disparity suggests that clustering criteria with high movement are likely more indicative of dynamic developmental patterns and socio-economic progress.

Clusters experiencing substantial changes in their member states provide critical insights into regions undergoing transformation, as they reflect impactful factors driving development. As noted in a study by Lin et al. (2013), "*Clusters with significant movement in their composition indicate dynamic socio-economic interactions and changing developmental landscapes, which can act as key markers for identifying regions of progress or decline*" [7]. Conversely, criteria with limited transitions may highlight entrenched stagnation or indicate less relevance to evolving developmental trajectories.

The Sankey diagrams and heatmaps also reveal the influence of spatial inter-dependencies. States transitioning from lower-performing to higher-performing clusters often share boundaries with developed regions, reinforcing the role of regional proximity in fostering socio-economic progress. Conversely, limited movement in some criteria highlights the persistence of systemic barriers in less responsive areas, necessitating tailored interventions to address these challenges. This underscores the importance of selecting clustering criteria that align with indicators reflecting dynamic regional development, as also suggested in prior research: "The choice of clustering features significantly impacts the interpretability of socio-economic patterns, especially in regions marked by disparities" [8].

5) Feedback: The second iteration revealed that greater cluster movement reflects the significance of the underlying feature in driving socio-economic change, consistent with findings in prior research [7]. To retain these insights, cluster transition data was saved to `iter2.csv` for analysis and refinement in the next iteration.

F. The Visual Analytics Workflow - Third Run

1) Data: Since from the previous iteration we knew that clusters don't equally influence whether a state is "Good" or "bad" (so the features influencing this also not equally determine), so we thought to combine these features weighted by some form cluster movements. We assigned weights to features based on the movements between clusters in the second iteration. The process involved the following steps :

a) Data Preparation: The datasets for 2001 and 2011 were combined, and the data was standardized to ensure consistency across features.
b) Weights were assigned to clustering criteria proportional to movement between clusters in that criteria.
c) PCA for Feature Loadings: PCA was applied to each clustering criterion's features, with the first component's absolute loadings normalized to sum to one, prioritizing features relevant to cluster transitions.
d) Weight Calculation: For each feature, the normalized PCA loading was multiplied by the corresponding clustering criterion's weight based on cluster movement.
e) Weighted Score: A new column, `Weighted_score`, was added, representing the weighted sum of the standardized feature values

2) ML Model: K-means clustering was applied to the feature `Weighted_score`. The number of clusters, $K = 3$, was determined using the method described by Andrew Ng [6].

3) Visualisation: Two 1D scatter plots were created to represent the clusters based on `Weighted_score` for the years 2001 and 2011 as shown in Figure 31. Additionally, a choropleth map was generated to visualize the cluster assignments for each state 32.

4) Knowledge:

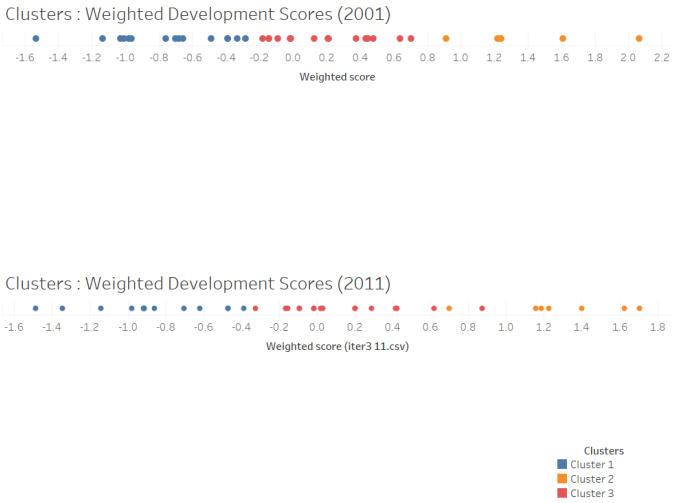

Fig. 31: Clustering on `Weighted_score` (2001 and 2011)

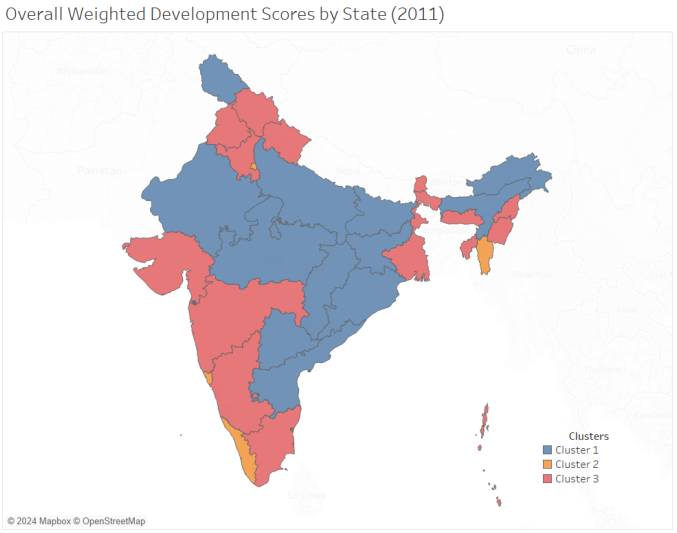

Fig. 32: State-wise Distribution of Overall Weighted Development Scores (2011): Clusters Representing Regional Development Patterns

a) Each state was categorized into one of three groups: - contributors (and rapid improvers), moderately improvers and Stagnantors (Low or nill improvers).
b) The clusters are not randomly distributed; rather, they exhibit spatial partitioning, suggesting that regional factors influence cluster membership.

G. Final Hypothesis

We conclude the section by presenting our final hypothesis:

1) Urbanization drives higher and above education.
2) Literacy reduces gender disparities (at least in education) 22.
3) Population density negatively impacts literacy growth.
4) No single feature alone determines outcomes; features must be combined with proper weights.
5) Development trajectories exist, with neighboring states influencing each other.
6) Union Territories (UTs) have a minimal impact due to their smaller size and unique socio-economic conditions.

III. Task 2: Analyzing Literacy and Socio-Economic Dynamics in Female-Headed Households Through Census Data Visual Analytics

The workflow process is illustrated in 33 and elaborated upon in the subsequent section. This assignment begins with a summary of Task 2 from Assignment 1, providing an overview of the visualization methodology and the inferences drawn from it. These visualizations serve as the foundation for deriving knowledge during the initial iteration of our workflow. Based on the insights gained, we identify specific areas within the methodology that offer opportunities for refinement. These improvements are subsequently integrated into the workflow in later iterations. The workflow depicted in 33 draws inspiration from the framework proposed by Kiem et al. [2].

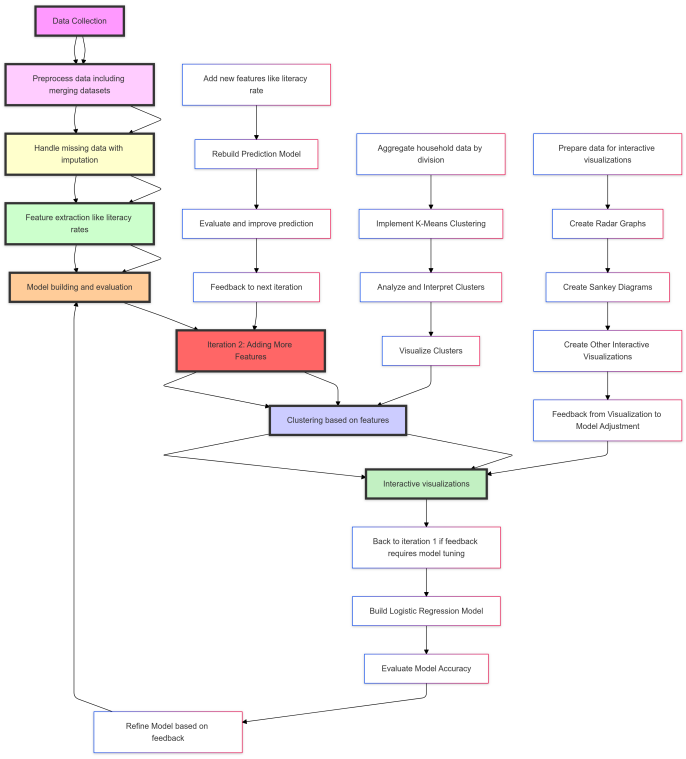

Fig. 33: Visual analytics workflow utilized in Task 2, illustrating iterative processes as described in Kiem et al. [2].

A. Visualizations and inferences from Assignment 1

This section serves as a summary of Task 2 from Assignment 1, providing readers with essential context before building upon its findings. The primary objective of the

task was to provide a comprehensive analysis of housing infrastructure and planning in India, focusing on the current state of urban and rural housing and challenges related to land use, affordability, and sustainable development. By visualizing various parameters, key insights were gained into household sizes and the effect of education on household structure.

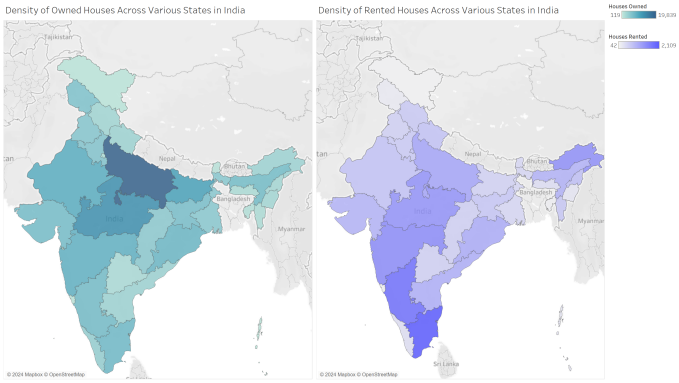

Fig. 34: Density of Owned and Rented Houses

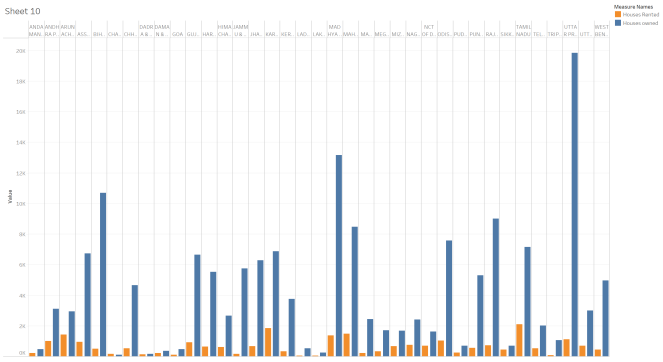

Fig. 35: Statewise owned and rented houses

Visualizations:

1) **Choropleth Maps:** Two chorpleth maps 34 and 35 visualized the statewise density of owned and rented households, highlighting distinct regional patterns. Northern states like Punjab, Uttar Pradesh, etc., have higher densities of owned houses while southern states like Karnataka, Tamil Nadu, etc., have higher densities of rented households.
2) **Grouped Bar Chart:** The Grouped Bar chart 36 visualized a very important trend of household sizes. This also highlighted distinct regional patterns. Northern states like Bihar, Uttar Pradesh, etc., have larger household sizes while southern states like Karnataka, Tamil Nadu, etc., are more inclined towards nuclear families.
3) **Stacked Bar Chart:** This visualization 37 also aimed to highlight the same pattern as intended by Grouped Bar Graph in 36.

Inference: The visualizations of household ownership and rental density, as well as household size patterns, provide

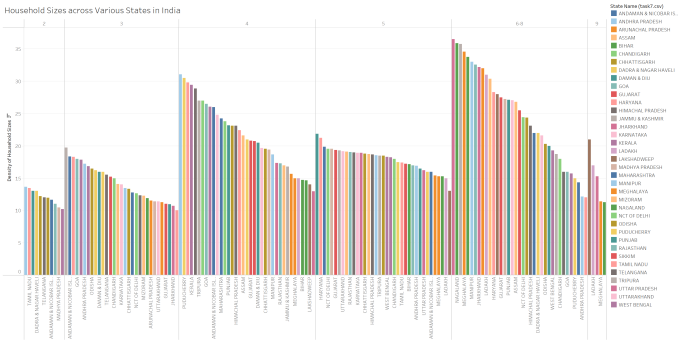

Fig. 36: Statewise Trend in Household Sizes in a Decreasing Fashion.

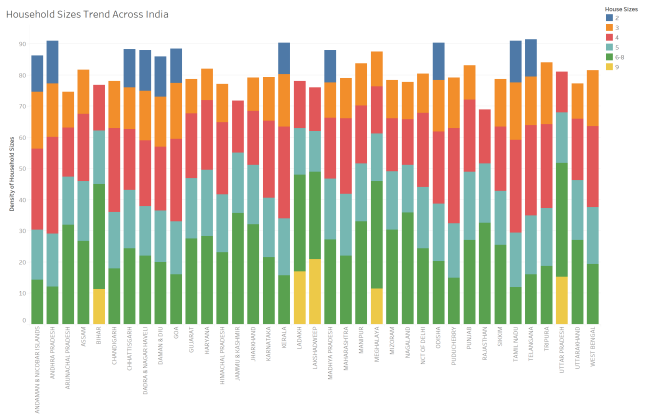

Fig. 37: Household sizes trend across India.

critical insights into the socio-economic dynamics of female-headed households. These findings can be interpreted as follows:

1) **Statewise Ownership and Rental Patterns (Choropleth Maps)**
   - **Owned Households:** States with a higher density of owned households, such as Punjab and Uttar Pradesh, may reflect cultural and economic structures that favor stable, long-term family setups. This stability could provide a more supportive environment for female-headed households, enabling them to access resources and maintain literacy rates.
   - **Rented Households:** Southern states like Karnataka and Tamil Nadu, which show a higher density of rented households, may reflect more urbanized and mobile populations. For female-headed households, this mobility could correlate with greater economic independence but also highlight potential vulnerabilities, such as instability in housing and access to education.
2) **Household Sizes (Grouped Bar Chart & Stacked Bar Chart)**
   - **Larger Households in Northern States:** The trend of larger household sizes in states like Bihar and

Uttar Pradesh suggests extended family living arrangements, which might provide additional social support for female heads of households. However, these arrangements may also increase the economic burden, potentially impacting literacy levels among dependents.

- **Nuclear Families in Southern States:** The inclination towards nuclear families in states like Karnataka and Tamil Nadu often reflects urbanization and socio-economic modernization. For female-headed households, this shift could signify greater autonomy and decision-making power, positively influencing literacy rates. However, the lack of extended family support could pose challenges in managing household dynamics.

B. Additional Dataset

The dataset utilized for Assignment 1 was sourced from [1], a subset of the official Indian Government's Census dataset [3]. The latter was selected for its inclusion of crucial columns, such as urbanization data, which were absent in the initial dataset. Furthermore, the initial dataset exhibited inconsistencies in calculations, such as discrepancies where the sum of literates, illiterates, and individuals who did not disclose information did not align with the total population. These issues rendered the initial dataset less reliable, prompting the adoption of the more comprehensive and accurate dataset from [3]. Also, the latter dataset [3] was taken directly from the government website, making it exhaustive and much more reliable.

C. Data Processing

1) **Missing Data Handling:** Replaced NaN values using mean for numerical and mode for categorical columns.
2) **Feature Engineering:** Created a `Literacy Rate` column:
$$\text{Literacy Rate} = \frac{\text{Number of Literates}}{\text{Total Heads}} \times 100$$
Ensured 0 was assigned where `Total Heads` was 0.
3) **Standardization and Normalization:** Standardized numerical features like `Literacy Rate`, `Number of Literates`, and household sizes using `StandardScaler`.
4) **Aggregation:** Grouped data by divisions or regions to calculate averages for literacy rates, female-headed households, and household sizes.
5) **Reshaping and Transformation:** Converted household size data (1, 2, 3, etc.) into a cumulative format for radar charts. Reformatted ownership and rental density data for statewise comparisons.
6) **Clustering:** Applied K-Means clustering to identify socio-economic patterns using optimal cluster determination methods. Assigned cluster labels for further analysis.
7) **Preparation for Visualizations:** Merged division-level metrics like literacy rate, female head percentage, and household sizes for advanced visualizations. Reformatted data for radar charts, Sankey diagrams, and bar plots.

Each step ensured a structured approach to analyzing socio-economic dynamics in female-headed households.

D. Visual Analytics Workflow

1) First Run:

a. **Data**

a) **Data Cleaning**

- Missing values and inconsistencies in the data were handled, ensuring all the required columns (such as state names, population counts, household ownership, and household sizes) were cleaned and structured appropriately.
- Any invalid entries, such as outliers or duplicates, were removed to avoid skewing the analysis.

b) **Data Transformation**

- The data was reshaped to ensure it could be used effectively for visualizations.
- Household size data was categorized into distinct columns for easy aggregation and analysis. Columns like household size categories (1, 2, 3, 4, 5, 6, and 7+) were created to capture the distribution of household sizes.

c) **Geospatial Data Handling**

- The census data was mapped to geographical regions, ensuring it could be represented on a map, especially for the choropleth visualization of female-headed households.

b. **Models**

In Iteration 1, a predictive model was developed to classify the ownership status of female-headed households based on socio-demographic features. The primary objective was to understand the factors influencing ownership decisions within these households. A Decision Tree Classifier was selected as the model due to its simplicity, interpretability, and ability to handle both categorical and numerical data. This section outlines the approach taken for model development, the rationale behind the choice of model, and an overview of the model's performance.

a) **Model Selection**

The Decision Tree Classifier was chosen for its capacity to model non-linear relationships between features while providing a transparent and interpretable structure. Decision trees are advantageous for categorical variables and can naturally handle both numerical and categorical data without requiring explicit transformations. The model works by recursively splitting the data based on feature values, creating a tree structure where each internal node represents a feature split, and each leaf node

represents a prediction outcome. This makes it especially valuable in this context, where understanding the relationship between socio-demographic features and household ownership is crucial.

The target variable in this model, `Ownership`, represents whether a household is owned or rented. The features used for prediction included `Division` (a categorical variable representing geographical regions) and `Household Size` (a numerical variable indicating the number of people in the household). These features were selected based on their potential influence on housing ownership, with the hypothesis that household size and geographical location may significantly impact the likelihood of ownership.

b) **Training the Model**

The model was trained using the socio-demographic data, where the target variable `Ownership` was mapped to the features `Division` and `Household Size`. The `Division` variable was encoded numerically to allow it to be used effectively by the model, while `Household Size` was directly utilized as a numerical input feature.

The model was then fit using the training data, and its predictions were made based on the encoded features. The Decision Tree Classifier was trained to identify patterns in the data that distinguish between owned and rented households. The tree-building process involves evaluating the best feature splits based on criteria such as Gini impurity or entropy, which quantify the "impurity" or uncertainty at each node of the tree.

c) **Prediction and Evaluation**

Once trained, the model was used to make predictions on new data points, with sample inputs such as `Division` and `Household Size` representing the socio-demographic characteristics of a household. The model outputs a classification (owned or rented) based on the values of the input features.

To assess the effectiveness of the model, further evaluation would typically involve using additional metrics such as accuracy, precision, recall, and F1-score. These metrics would provide a more complete picture of how well the model performs in distinguishing between owned and rented households. Cross-validation could also be employed to evaluate the model's generalizability and reduce the risk of overfitting.

c. **Visualizations**

a) **Choropleth Map**

Chloropleth Map in 38 has the following properties:

- **Data Inputs:** States and the percentage of female-headed households.

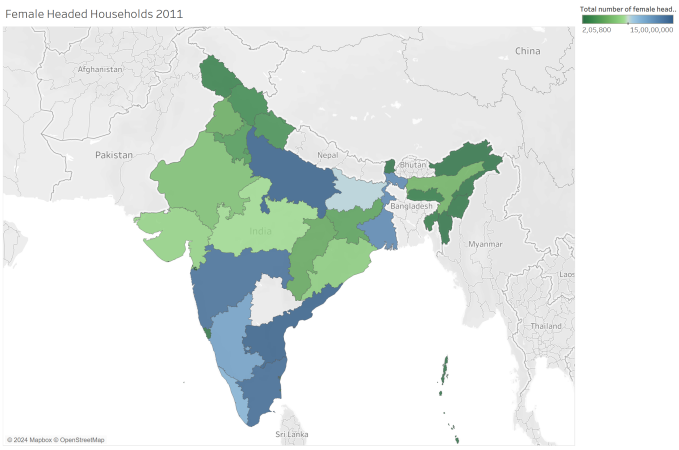

Fig. 38: Density of Female Headed Households in 2011

- **Color Palette:** A color gradient was applied to represent varying levels of female-headed households, where darker shades indicated higher percentages.
- **Purpose:** The map helped in identifying regions with higher or lower densities of female-headed households, contributing to understanding regional socio-economic differences.

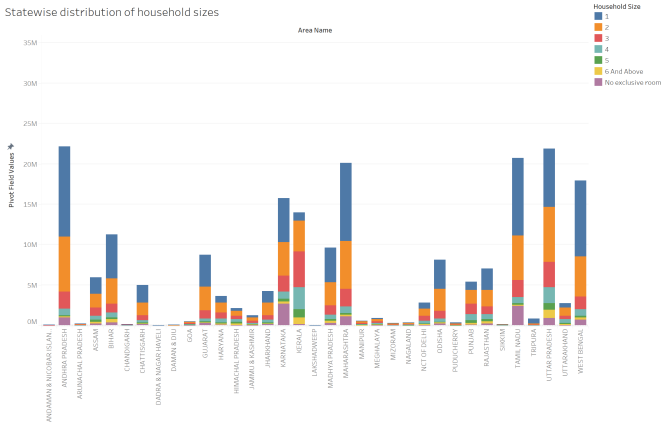

Fig. 39: Distribution of different Household Sizes across States

b) **Grouped Bar Chart**

Grouped Bar Chart in 39 has the following properties:

- **Data Inputs:** State-level household size distributions.
- **Colors:** Different colors were used to represent each household size category, enabling easy identification of household composition trends across regions.
- **Inferences:**
  - The bar chart highlighted regions where larger household sizes were predominant (e.g., northern states like Bihar, Uttar Pradesh).

– Southern states (e.g., Tamil Nadu, Karnataka) showed a preference for nuclear families.

d. **Knowledge**

The visualizations presented in Iteration 1 provided a comprehensive view of the socio-economic dynamics surrounding female-headed households:

a) **Regional Patterns in Female-Headed Households**

- States with higher female-headed household densities were identified, highlighting areas where socio-economic factors such as gender dynamics, employment patterns, and migration might influence family structures.

b) **Household Size Trends**

- Larger households were predominantly found in northern states (e.g., Bihar, Uttar Pradesh), suggesting the continuation of extended family living arrangements. This could correlate with support systems available to female heads of households.
- Southern states, which favored nuclear families (e.g., Tamil Nadu, Karnataka), indicated a shift toward more autonomous family structures that may provide better opportunities for female-headed households in terms of decision-making and financial independence.

e. **Feedback**

The feedback for Iteration 2 emphasized the need for deeper insights and a more robust predictive framework. While Iteration 1 introduced basic modeling with a decision tree classifier, it lacked socio-economic context and relied on limited variables, reducing its interpretability. Visualizations, though informative, were insufficient to capture regional and categorical variations. The feedback led to expanding the feature set to include socio-economic factors, enhancing visualizations with geospatial and categorical analyses, and refining the modeling approach to a Random Forest Classifier for improved accuracy and generalizability. This feedback significantly enhanced the depth and clarity of the analysis in Iteration 2.

2) Second Run:

a. **Data**

Data preprocessing forms the foundation of any data-driven analysis. In this research, data was primarily drawn from the 2011 Census of India and included attributes such as household sizes, female and male-headed households, and literacy rates. The following preprocessing steps were implemented to ensure a clean and analyzable dataset:

a) **Data Cleaning:**

- Columns containing missing or inconsistent entries were addressed by using imputation techniques or removing invalid entries to prevent biases during analysis. For instance, numeric columns such as Number of Literates were transformed using coercion and imputation.
- Extreme values in attributes like household sizes and literacy rates were identified and appropriately managed to avoid skewed patterns.
- Variables were converted into the appropriate data types (e.g., converting household sizes from strings to integers).

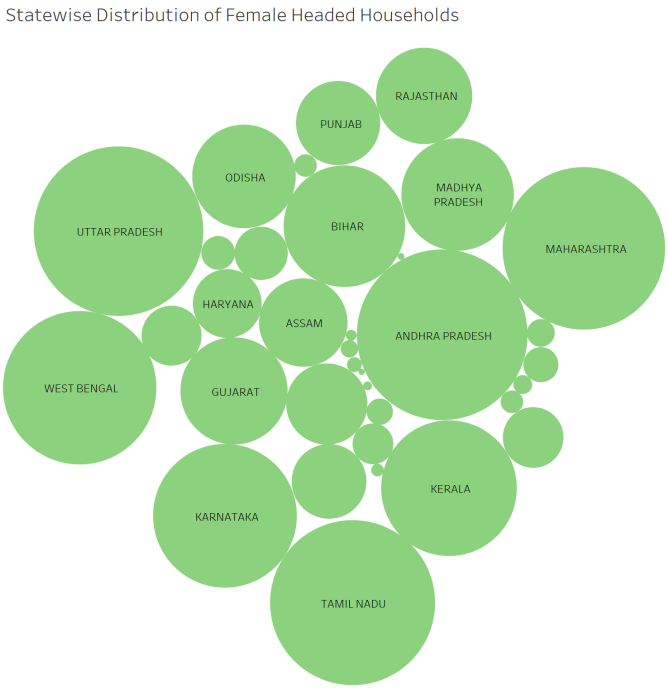

Fig. 40: Statewise distribution of Female Headed Households

b) **Data Transformation:**

- Household size data was normalized to represent the percentage distribution of each category (e.g., 1-person to 7+ persons) within the total number of households. This enabled consistent comparisons across states.
- New variables were created, such as the literacy rate, calculated as the ratio of literates to total household heads, and regions were encoded numerically for analysis.
- The variable Religion was encoded numerically for machine learning compatibility.

c) **Data Aggregation**

- The data was grouped by divisions (regions) to calculate average values for attributes such as literacy rate, male and female household heads, and household size percentages. Aggregation facilitated insights at both macro and micro levels.

b. **Model**

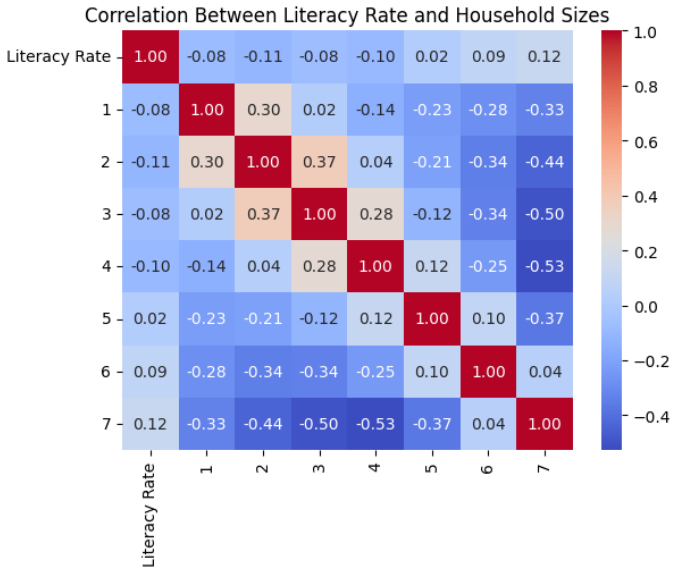

Fig. 41: Correlation between literacy rate and household size

While the primary objective was exploratory visualization, logistic regression modeling was conducted to predict the likelihood of a household being female-headed.

a) **Overview**

- **Features:** Predictors included the literacy rate, religion (encoded numerically), and household size distributions.
- **Target:** The target variable was a binary indicator of whether a household was female-headed (1) or male-headed (0).
- **Data Split and Imputation:** Data was split into training (80%) and testing (20%) sets. Missing data in numeric columns was imputed with the mean strategy.
- **Algorithm:** Logistic regression was chosen due to its simplicity and interpretability for classification tasks.

b) **Evaluation**

- Performance was measured using metrics like precision, recall, and F1-score. Results highlighted the key predictors of female-headed households, with literacy rate and household size distribution emerging as significant factors.

c. **Visualization**

The data analysis was complemented by high-quality visualizations, enabling intuitive and effective exploration of patterns and trends:

- **Bubble Chart:** The visualization 40 depicted the distribution of female-headed households across states. Bubble sizes represented the absolute count of female-headed households, while the geographic arrangement facilitated comparisons between states. Larger bubbles in states like Uttar Pradesh, Bihar, and Maharashtra indicated higher densities.
- **Correlation Heatmap:** A heatmap 41 was used to visualize correlations between literacy rates and household sizes. Notable observations included a negative correlation between literacy rates and larger household sizes (e.g., 7+ person households), highlighting an inverse relationship between literacy and extended family structures.

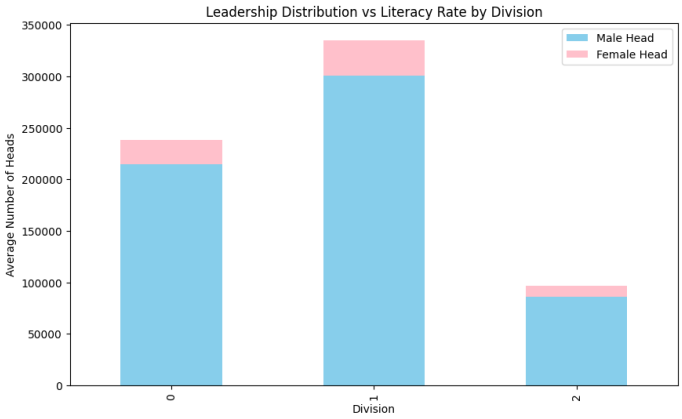

Fig. 42: Leadership Distribution vs Literacy Rate by Division

- **Stacked Bar Chart:** This chart 42 compared the average number of male and female heads across divisions, normalized by literacy rates. A significant observation was the dominance of male heads in regions with lower literacy rates, whereas higher literacy correlated with a greater proportion of female-headed households.

d. **Knowledge**

The analysis revealed several socio-economic and demographic insights into the dynamics of female-headed households in India:

- **Regional Trends**
  - Northern states such as Uttar Pradesh and Bihar exhibited larger households and a lower proportion of female-headed households. This aligns with traditional family structures favoring extended families and male leadership.
  - Southern states, such as Tamil Nadu and Kerala, demonstrated smaller household sizes and higher literacy rates, correlating with a higher prevalence of female-headed households. This suggests evolving family structures and improved socio-economic opportunities for women.
- **Correlation Between Literacy and Household Dynamics**
  - Higher literacy rates were associated with smaller household sizes and a higher probability of female-headed households. This highlights the role of education in empowering women and fostering nuclear family systems.

- **Implications for Policy**
  - States with higher literacy rates and smaller households offer insights into pathways for improving socio-economic outcomes for women. Investments in education and targeted support for female-headed households in underdeveloped regions could mitigate disparities.

e. **Feedback**

This iteration of data processing, modeling, and visualization highlights critical socio-economic patterns but also reveals areas for refinement to enhance insights in future iterations. For the next iteration, incorporating more granular data, such as rural-urban distinctions and caste-based segmentation, could uncover nuanced disparities within female-headed households. Refining the model by exploring non-linear algorithms or ensemble methods may better capture complex relationships between variables like literacy rate, household size, and female leadership. Additionally, visualizations can be improved by integrating interactive elements, such as drill-down capabilities or regional comparisons over time, to facilitate deeper engagement and discovery. These enhancements would not only strengthen the analytical rigor but also provide actionable insights for policy-making.

3) Third Run:

a. **Data**

- **Standardization:** The dataset contained numerical variables such as literacy rates, total heads, and the number of literates, each with different scales. To ensure that the clustering algorithm effectively measures distances between data points, standardization was applied using the StandardScaler. This transformation scaled the features to a mean of zero and a standard deviation of one.
- **Feature Engineering**
  - **Aggregation of Household Sizes:** Household size categories were summed to compute a total household size for each division, providing a holistic view of household distributions.
  - **Percentage Computation:** The percentage of each household size relative to the total was calculated. This metric helped contextualize the prevalence of specific household sizes across divisions.

b. **Model**

- **Elbow Method:** To determine the optimal number of clusters, the Elbow Method was employed as in 43. The distortions (sum of squared distances from cluster centroids) were plotted against the number of clusters. A noticeable "elbow point" indicated the ideal number of clusters, balancing between underfitting and overfitting.
- **Silhouette Score:** Silhouette analysis was conducted for cluster validation. This metric evaluates how similar a point is to its cluster compared to other clusters. The optimal number of clusters was identified as the one yielding the highest silhouette score.
- **Hierarchical Clustering:** Hierarchical clustering using the Ward linkage method was performed, producing a dendrogram that illustrated the merging process of data points as in 44. This provided an alternative perspective to the K-Means algorithm and confirmed the stability of the clustering results.

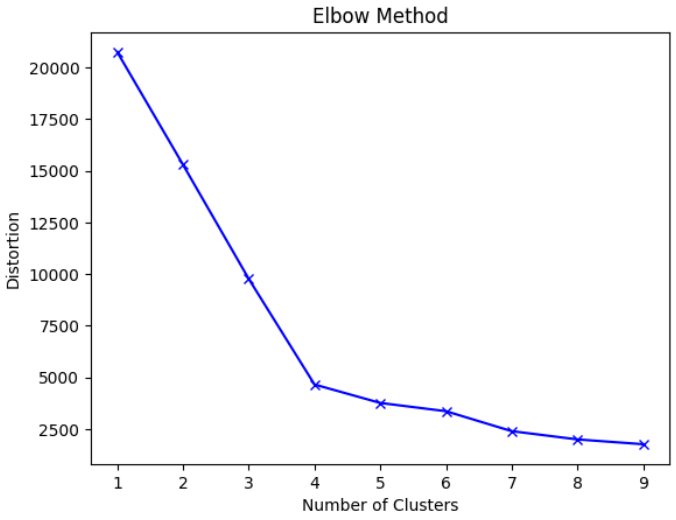

Fig. 43: Elbow Method

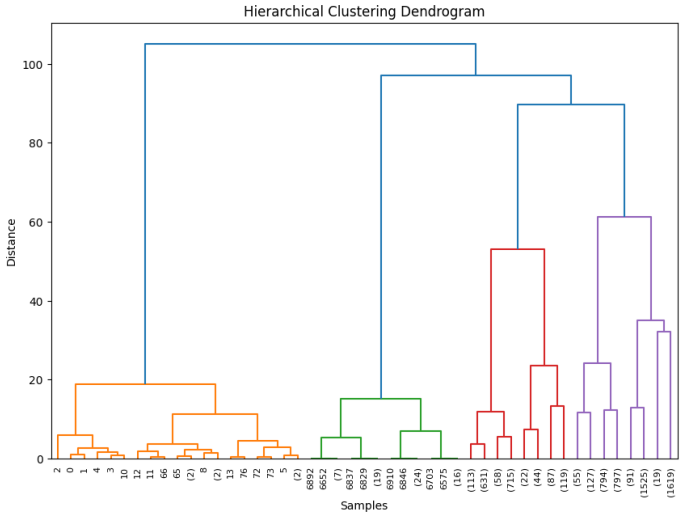

Fig. 44: Dendogram

c. **Visualizations**

- **Radar Chart:** The radar chart as shown in 45 visualized household size distributions across divisions (Rural, Urban, and Total). Key observations include:
  - Urban and rural areas exhibit significant variation in household size distribution, reflecting differ-

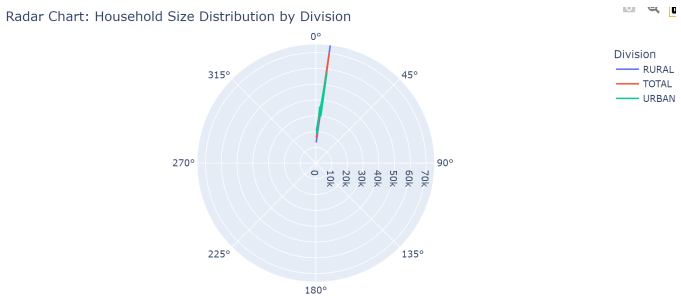

Fig. 45: Household Size Distribution by Division

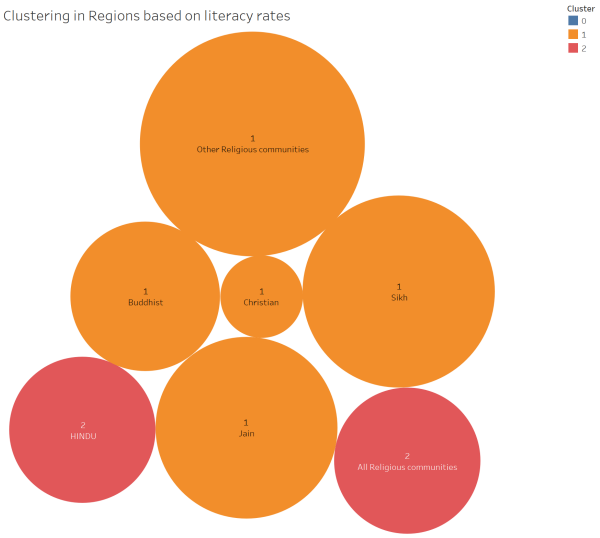

Fig. 46: Clustering in Regions based on literacy rates

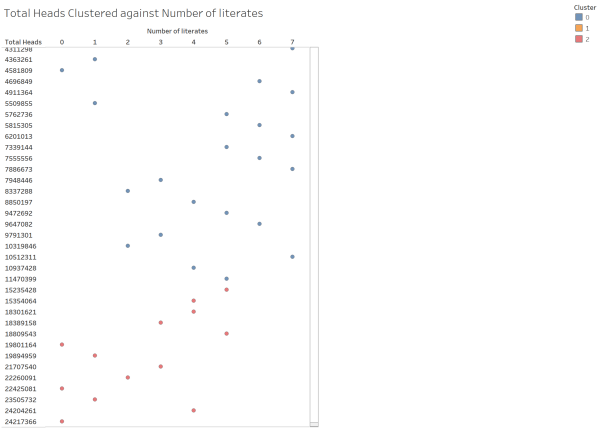

Fig. 47: Total Heads Clustered against Number of literates

ences in living arrangements.
  - Total distributions closely mirror a composite of rural and urban distributions.

  This visualization provided an intuitive comparison of household patterns across different divisions, enabling policymakers to identify regional disparities.
- **Bubble Chart:** A bubble chart as show in 46 was used to illustrate clusters formed based on literacy rates for various religious communities. Key findings include:
  - Cluster 1, primarily representing smaller religious communities, exhibited distinct literacy characteristics.
  - Cluster 2, comprising larger communities like Hindus, demonstrated different trends in literacy, likely influenced by population size and socio-economic factors.

  The bubble sizes corresponded to population sizes, adding a multidimensional aspect to the visualization.
- **Scatterplot:** This scatterplot as shown in 47 explored the relationship between total heads (population size) and the number of literates, with clusters distinguished by colors. Notable observations:
  - Cluster 0 primarily represented low-literate populations.
  - Higher-literate populations belonged to distinct clusters, suggesting varied literacy patterns based on demographic factors.

d. **Knowledge**
- **Socio-Economic Insights**
  - Clustering highlighted disparities in literacy rates across regions and communities. This underscores the need for targeted educational policies.
  - Household size distributions revealed fundamental differences between urban and rural living arrangements, informing housing and infrastructure development.
- **Analytical Strengths**
  - Standardization ensured robust clustering performance, avoiding bias from feature scaling.
  - The combination of visualizations provided complementary perspectives, enriching the analysis and facilitating stakeholder understanding.
- **Challenges and Opportunities**
  - The presence of overlapping clusters in some visualizations suggested the need for additional features to enhance model discrimination.
  - Hierarchical clustering, while insightful, was computationally expensive for larger datasets, indicating scalability issues.

e. **Feedback**

The next iteration should focus on refining data processing by introducing domain-specific transformations,

handling missing values with advanced imputation, and exploring feature interactions. Alternative clustering methods like DBSCAN or Gaussian Mixture Models should be tested to enhance robustness, alongside validation metrics to assess stability. Visualizations can be improved with interactive tools and multidimensional approaches, such as t-SNE or geospatial mapping, to make patterns more intuitive. Incorporating domain expertise and stakeholder feedback will ensure insights are actionable, while hypothesis-driven exploration and enhanced cluster profiling will deepen understanding and guide future analyses.

E. Summary

Our analysis focused on understanding literacy and socio-economic dynamics in female-headed households. We used visualizations, regression and clustering methods to learn more about these differences.

- **First Iteration:** Focused on data cleaning, addressing missing values, and mapping census data to regions. Household sizes were categorized, and a Decision Tree Classifier was used to predict ownership of female-headed households. Socio-demographic attributes like division and household size were key features.
- **Second Iteration:** Incorporated clustering for analyzing household attributes and their relationships with socio-economic indicators. Used K-Means to group households and identified patterns in family composition and ownership trends. Results guided the creation of visualizations highlighting regional ownership disparities.
- **Third Iteration:** Emphasized interactive visualizations with radar charts, Sankey diagrams, and geospatial heatmaps. Focus shifted from modeling to visual storytelling to explore trends in gender, household size, and ownership. This approach provided actionable insights into socio-economic dynamics.

IV. Task 3: Analyzing Workforce Demographics in India.

The workflow process is illustrated in 51 and elaborated upon in the subsequent section. This assignment begins with a summary of Task 3 from Assignment 1, providing an overview of the visualization methodology and the inferences drawn from it. These visualizations serve as the foundation for deriving knowledge during the initial iteration of our workflow. Based on the insights gained, we identify specific areas within the methodology that offer opportunities for refinement. These improvements are subsequently integrated into the workflow in later iterations. The workflow depicted in 51 draws inspiration from the framework proposed by Kiem et al. [2].

A. Visualizations and inferences from Assignment 1

This section serves as a summary of Task 3 from Assignment 1, providing readers with essential context before building upon its findings. The primary objective of the task was to analyze workforce demographics and distribution across India. Through this analysis, key insights were gained into gender composition, employment categories, and regional workforce participation patterns, laying the foundation for further exploration in this assignment.

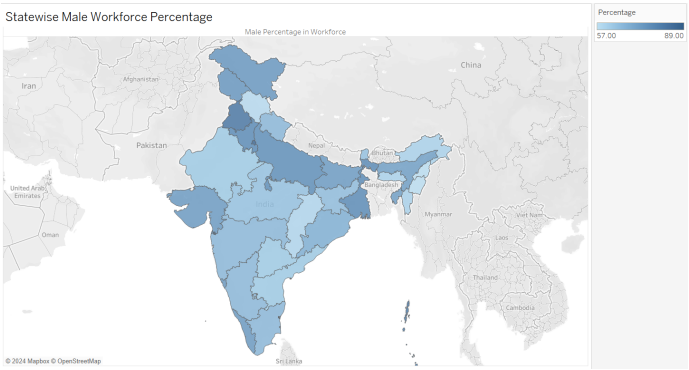

Fig. 48: State-wise male workforce percentage derived from Assignment 1.

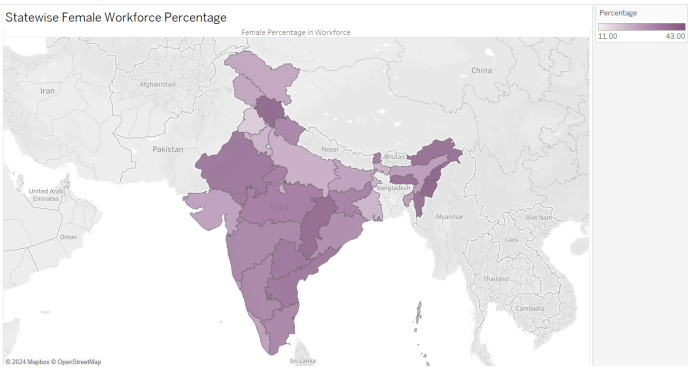

Fig. 49: State-wise female workforce percentage derived from Assignment 1.

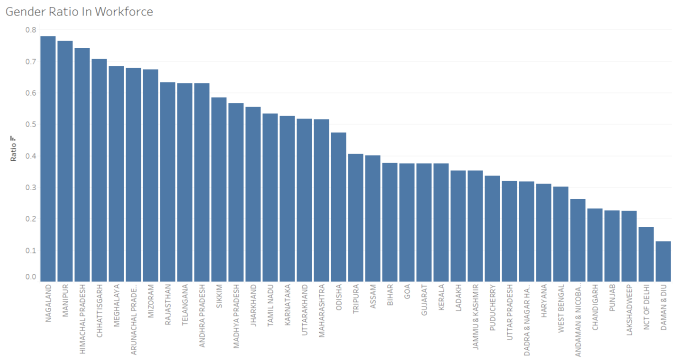

Fig. 50: State-wise gender ratio in the workforce, as analyzed in Assignment 1.

Visualizations:

1) **Choropleth Maps:** Two choropleth maps visualized the percentage of male and female workforce participation across Indian states 49 and 48, highlighting distinct

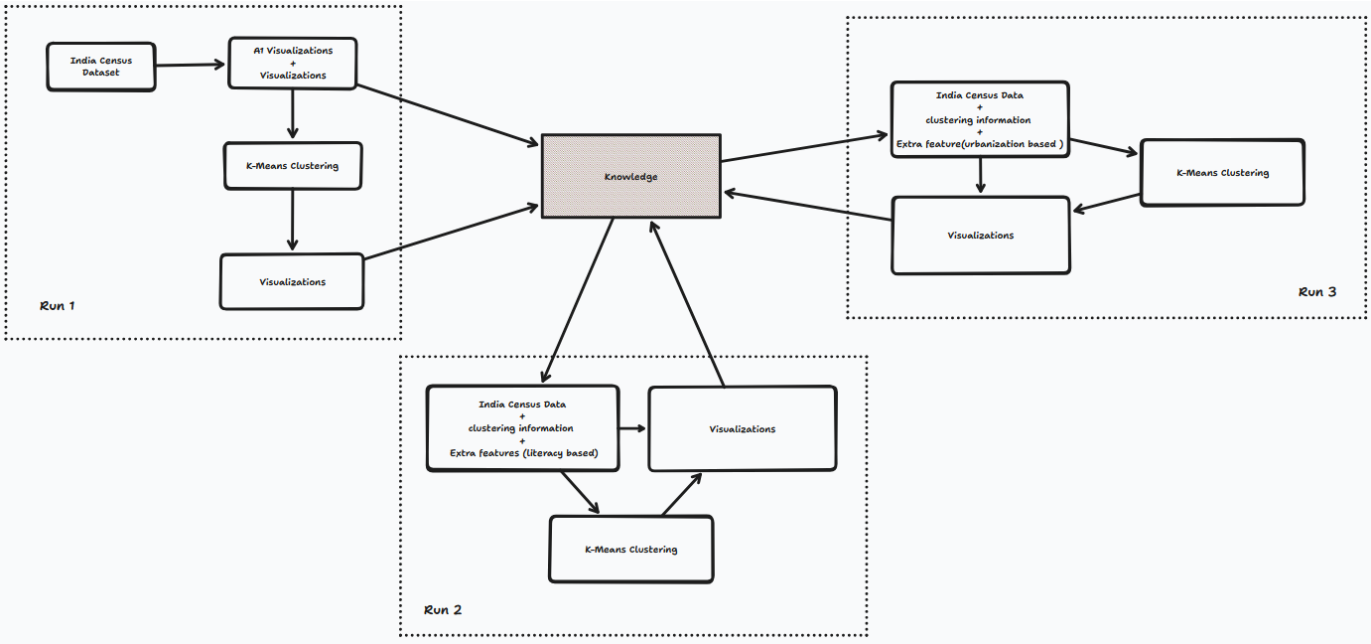

Fig. 51: Visual analytics workflow utilized in Task 3, illustrating iterative processes as described in Kiem et al. [2]. Each rectangle with dotted side represents a complete workflow cycle.

regional patterns. Northeastern states like Nagaland and Manipur show higher female participation, while northern and western states such as Punjab and Rajasthan exhibit lower percentages.

2) **Bar Chart on Gender Ratios:** A bar chart 50 illustrated gender workforce participation ratios, showcasing states like Himachal Pradesh and Chhattisgarh with high gender parity, while union territories like Daman and Diu displayed low ratios.
3) **Pie Chart of Worker Types:** A pie chart categorized the workforce into main, marginal, and non-workers. Female participation trends are notably higher in the marginal and non-worker categories.
4) **Bubble Chart of Worker Distribution:** Statewise bubble charts highlighted regional differences in workforce composition, with states like Uttar Pradesh showing significant populations of non-workers, often reflecting educational or cultural factors.
5) **Stacked Bar Chart on Worker Categories:** A stacked bar chart illustrated employment types (agricultural, household, and other workers), emphasizing states like Maharashtra and Tamil Nadu's urbanized workforce dominated by "Other Workers."

Inferences:

1) **Female Workforce Participation:** Higher female workforce participation is observed in agricultural economies like Nagaland and Manipur, while urbanized states tend to show lower female engagement due to economic and societal constraints.
2) **Impact of Literacy:** States with higher literacy rates, such as Kerala, often exhibit better gender workforce ratios, indicating a positive correlation between education and workforce inclusion.
3) **Urbanization and Employment Shifts:** Urbanized and industrialized states, such as Maharashtra and Tamil Nadu, demonstrate a significant shift toward non-agricultural employment, highlighting economic diversification.
4) **Gender Gaps in Workforce Participation:** Northern states with lower female participation, such as Rajasthan and Uttar Pradesh, point to cultural and socio-economic barriers, including lower literacy and fewer employment opportunities for women.
5) **Education and Employment Balance:** Northeastern states with a relatively high female workforce often reflect a balance between education levels and traditional employment in agriculture, contrasting with the challenges faced by women in urban centers.

B. Additional Dataset

The dataset utilized for Assignment 1 was sourced from [1], a subset of the official Indian Government's Census dataset [3]. The latter was selected for its inclusion of crucial columns, such as urbanization data, which were absent in the initial dataset. Furthermore, the initial dataset exhibited inconsistencies in calculations, such as discrepancies where the sum of literates, illiterates, and individuals who did not disclose information did not align with the total population. These issues rendered the initial dataset less reliable, prompting the adoption of the more comprehensive and accurate dataset from [3].

C. Data Processing

The additional dataset provided data in the form of multiple tables, with each table corresponding to a specific component of the Census. To analyze the data from multiple perspectives and gain deeper insights, it was necessary to combine these tables. For efficient table handling, we used Microsoft Excel and Python's Pandas library.

The dataset was organized at the district level, along with the corresponding state information. As this assignment required state-level analysis, we aggregated the data state-wise.

1) Key Processing Steps:

- **Handling Missing Values:** As expected from a Census dataset, there were no null values. However, the column names were technical and specific to Census terminology. To improve readability and usability, we renamed the columns to more descriptive names.
- **Aggregation:** The dataset, initially structured at the district level, was aggregated to the state level. This was essential for state-wise analysis and visualizations.
- **Feature Engineering:** We created new columns derived from existing data to facilitate a more comprehensive analysis. For example:
  - **Urbanization Index:** Calculated as the ratio of the population living in urban areas to the total population (`Urban Population / Total Population`).

These steps ensured the dataset was prepared and structured effectively for analysis in this assignment.

D. The Visual Analytics Workflow

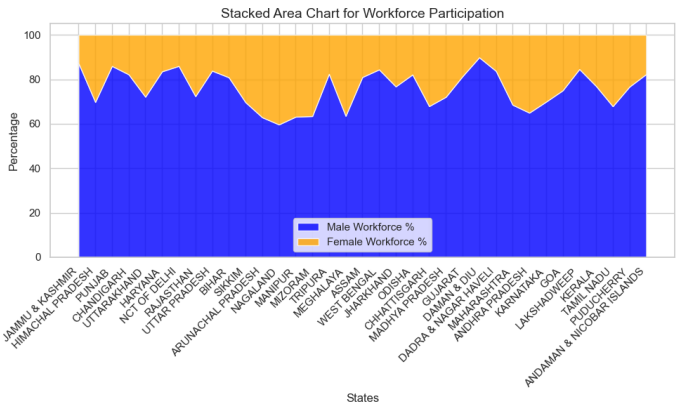

Fig. 52: Stacked area chart depicting the workforce participation of male and female workers.

1) The First Run:

a. **Data**
Data was used with minimal changes. We utilized columns such as the female workforce count and male workforce count, both state-wise, and created new columns representing each as a ratio of the total male and female workforce. This transformation allowed us to generate more meaningful visualizations and derive better inferences from the dataset.

b. **Models**
K-Means clustering was used to find natural groupings in the dataset. The main features we focused on for clustering were the male and female workforce participation rates in each state. By applying this method, we were able to group states based on their workforce demographics. The details of how K-Means was implemented, including the steps and results, are available in the submitted Jupyter notebooks. Additionally, the scatter plots before (53) and after (54) clustering are included in this report to help visualize the changes.

c. **Visualizations**

- Choropleth maps of male and female workforce percentages (48) and (49) were created to provide a better understanding of gender disparity in the workforce across states. These maps illustrate how the distribution of male and female workers varies by state.

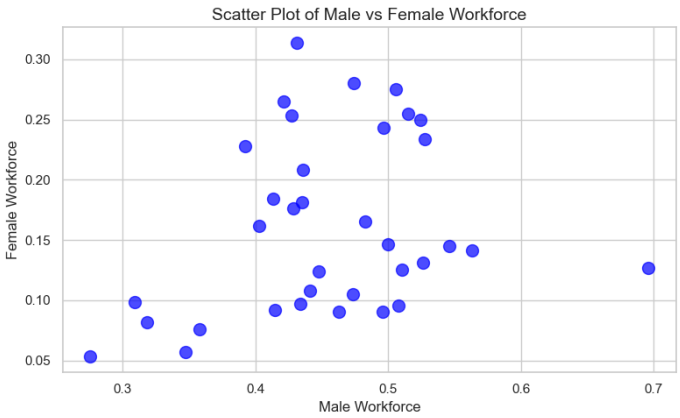

Fig. 53: Scatter plot of male and female workforce participation across different states before clustering.

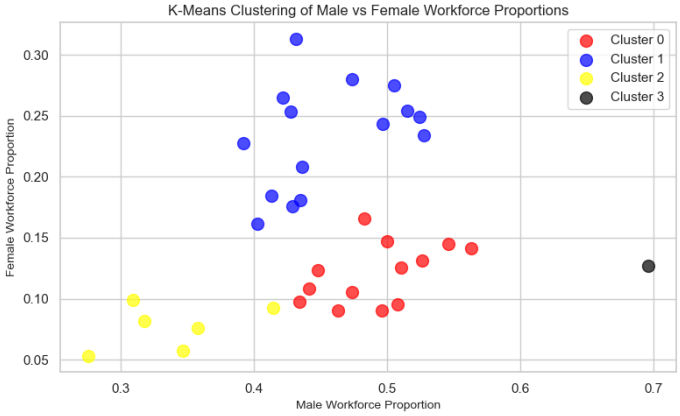

Fig. 54: Scatter plot of male and female workforce participation across different states with clusters marked after clustering.

- A bar chart showing the gender ratio in the workforce of each state (50) was included to highlight the relative disparity between male and female participation in the workforce. This visualization enables an easy comparison of gender ratios across states.
- A stacked area chart (52) of male and female workforce participation was used to display overall trends in workforce composition, both at the state level and for India as a whole. This helps visualize how male and female participation has evolved over time or across regions.
- Scatter plots of male and female workforce participation across states, before and after clustering (53) and (54), were generated to better understand the grouping of states based on their workforce demographics. These plots visually demonstrate how states were clustered based on their gender participation.

d. **Knowledge**

The choropleth maps 48 and 49 illustrates state-wise male and female workforce percentages, where darker shades represent higher participation. Northern and central states generally exhibit higher male workforce proportions, while some northeastern states appear lighter, indicating slightly lower male participation and higher female participation. This spatial visualization effectively highlights regional disparities in male workforce involvement.

The bar chart on gender ratio 50 in the workforce highlights the most equitable states, such as Nagaland and Manipur, where the female workforce ratio approaches parity with males. Conversely, states like Daman & Diu and NCT of Delhi show very low female workforce ratios. This chart underscores the significant variations in gender representation across states and identifies regions requiring targeted interventions for workforce gender equity.

The stacked area chart 52 reveals state-wise trends in male and female workforce participation. Across most states, male workforce percentages dominate, with female participation forming a relatively small fraction. Certain states show a slightly higher proportion of female workforce participation, leading to noticeable dips in the male workforce area. The chart emphasizes the persistent gender gap in workforce distribution on a state-by-state basis.

The unclustered scatter plot 53 of male vs. female workforce proportions shows significant variability between states, with male workforce participation generally higher. There is a noticeable range in female workforce proportions, with some states displaying higher gender equity while others have minimal female workforce presence. Without clustering, the pattern of gender-based workforce disparity is apparent but less structured.

The scatter plot showcasing K-Means clustering 54 of male vs. female workforce proportions reveals four distinct clusters. Cluster 0 (red) represents states with relatively high male workforce proportions and moderate female proportions. Cluster 1 (blue) indicates states where both male and female workforce proportions are relatively balanced. Cluster 2 (yellow) highlights areas with a low proportion of male workforce and very low female participation. Finally, Cluster 3 (black) is an outlier, representing a state with an unusually high male workforce proportion and low female representation. This clustering shows clear regional and proportional disparities in workforce participation by gender.

e. **Feedback**

We observed from the plots of literacy, which are presented in the subsequent run, that including columns related to literacy would provide a more comprehensive understanding of the data alongside the existing clusters. To address this, we incorporated additional columns for female literacy and cluster data from this run into the dataset. This enhancement allows for a deeper insight into the factors contributing to gender disparity in workforce participation.

We also added a new column, 'Weighted Female Workforce Participation,' which is calculated by multiplying female workforce participation by the female literacy rate. This provides a more comprehensive measure of gender disparity in workforce participation.

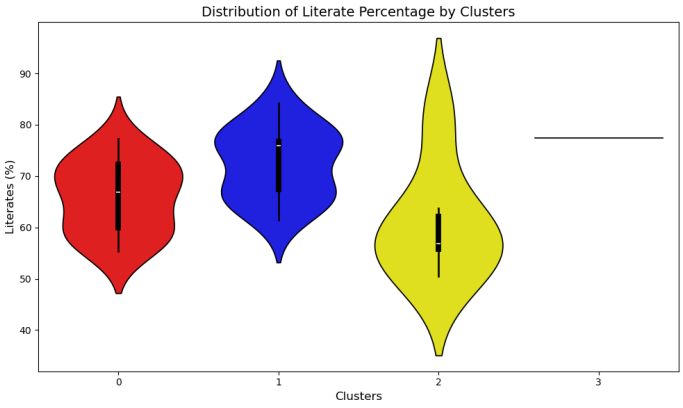

Fig. 55: Violin plot showing the distribution of literacy percentage, categorized by cluster.

2) The Second Run:

a. **Data**

The dataset now includes the columns from the previous run, along with several additional columns. These include the 'Cluster' from the previous clustering, 'Female Literacy Count' per state, 'Female Literacy Percentage,' and 'Female Literacy Ratio,' which are statistics derived from the female literacy count and total female population. Additionally, we introduced the 'Weighted Female Workforce Participation,' as described in the feedback section of the previous run.

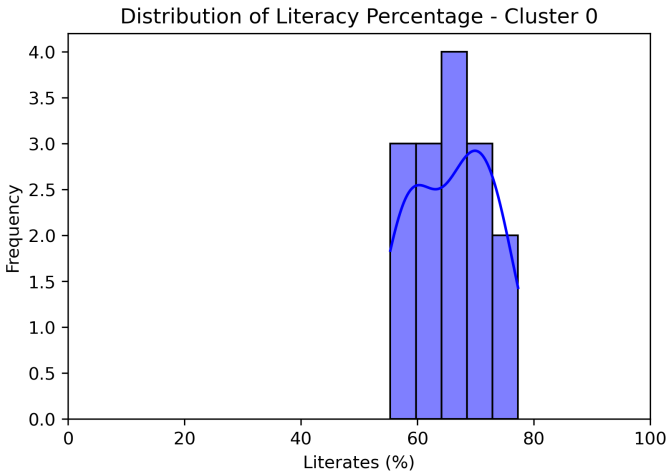

Fig. 56: Histogram with KDE for Cluster 0, showing the distribution of literacy percentage.

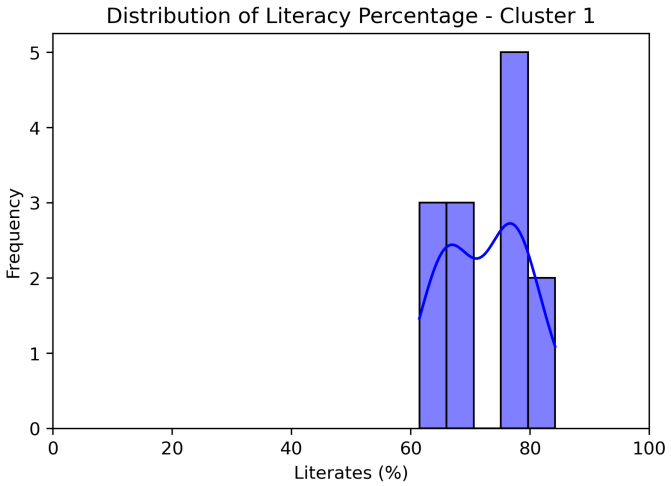

Fig. 57: Histogram with KDE for Cluster 1, showing the distribution of literacy percentage.

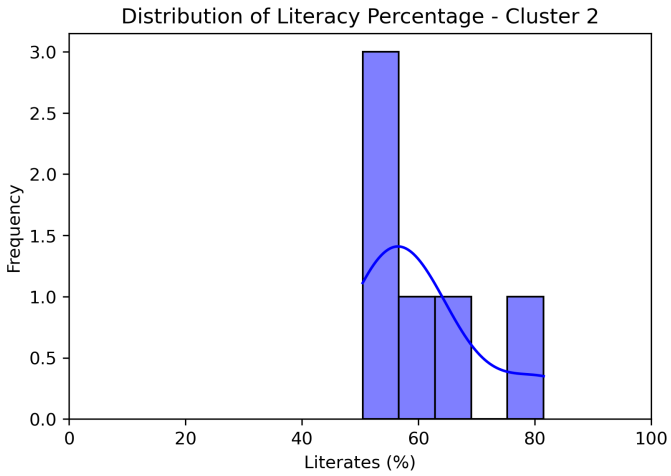

Fig. 58: Histogram with KDE for Cluster 2, showing the distribution of literacy percentage.

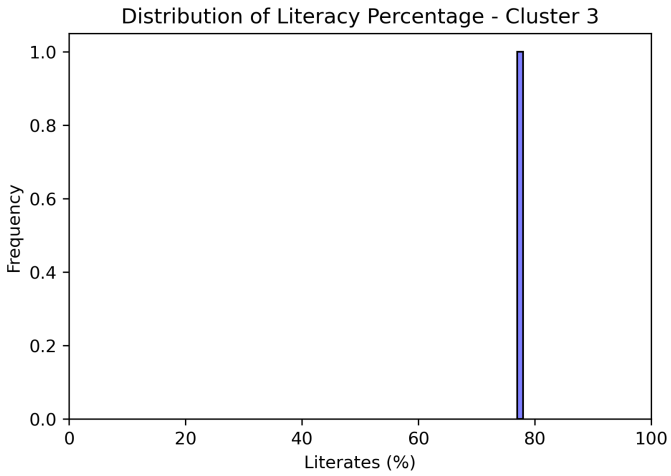

Fig. 59: Histogram with KDE for Cluster 3, showing the distribution of literacy percentage.

b. **Models**
We performed K-Means clustering again, this time experimenting with hierarchical clustering to explore any potential relationships with the previous clusters. However, we did not obtain any meaningful insights from the hierarchical clustering. Subsequently, we proceeded with K-Means clustering using the same features as the previous clustering, namely male workforce participation, female workforce participation, along with the newly added column for weighted female workforce participation. Additionally, the scatter plots before (53) and after (61) clustering are included in this report to help visualize the changes.

c. **Visualizations**

- Cluster-wise violin plots (55) were used to analyze the distribution of literates within each cluster, providing insights into why certain clusters perform better in terms of literacy.
- Histograms for each cluster (56, 57, 58, 59) were plotted to show the frequency distribution of literacy percentages across states.
- A scatter bubble chart (60) was created to visualize the newly added column, 'weighted female workforce participation,' for each state. The chart displays male workforce participation on the X-axis and female workforce participation on the Y-axis.
- A scatter plot (61) was generated after clustering based on male workforce participation, female workforce participation, and the new column, 'weighted female workforce participation,' to provide deeper insights into the clustering patterns.

d. **Knowledge**
The violin plot 55 combines the literacy percentage

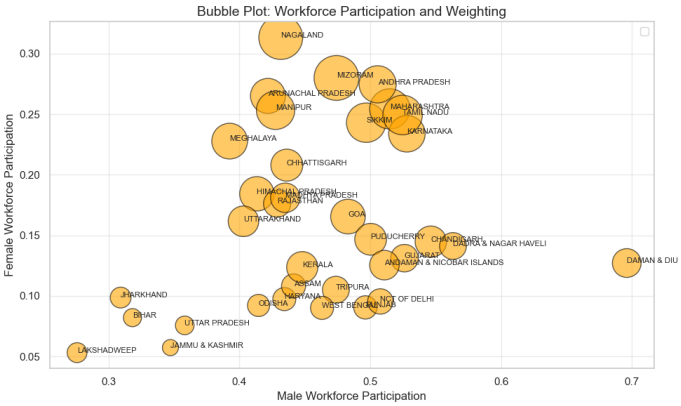

Fig. 60: Bubble chart representing male workforce participation on the x-axis, female workforce participation on the y-axis, and the bubble size corresponding to weighted female workforce participation (weighted by female literacy percentage).

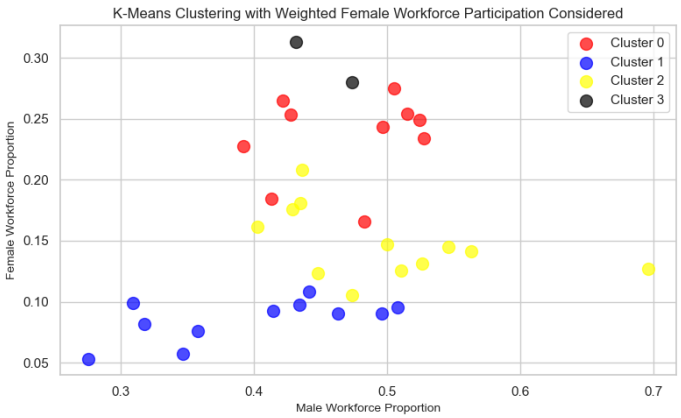

Fig. 61: Scatter plot of male and female workforce participation with clusters marked after clustering, using weighted female workforce as an additional parameter.

distributions of all clusters. Cluster 0 (red) demonstrates a moderately wide range of literacy percentages with the majority centered around 65-75%. Cluster 1 (blue) shows a slightly higher concentration and range of literacy percentages. Cluster 2 (yellow) displays the lowest literacy rates, with most states falling between 55% and 65%. Finally, Cluster 3 (black) stands out as an outlier with a consistently high literacy rate, reflecting the uniformity and unique workforce dynamics of this cluster. This shows that higher literacy rate means a higher female workforce particiaption.

From Histogram for cluster-0 56 we see that, most states in cluster 0 have literacy percentages concentrated between 60% and 80%, with a peak around 70%. Cluster 0 represents states with relatively high male workforce proportions and moderate female participation, which corresponds to the moderate and diverse literacy levels seen here. The histogram for Cluster-1 57 shows a more balanced and broader distribution of literacy percentages, ranging from approximately 60% to 80%. The data indicates relatively balanced literacy rates across states. Cluster 1 represents states where male and female workforce proportions are relatively balanced, possibly explaining the intermediate literacy rates seen in this cluster. The histogram for Cluster-2 58 displays a left-skewed distribution with most states having literacy percentages between 55% and 65%, and a few extending towards higher literacy levels (up to 80%). This indicates a moderate literacy range but lower than Cluster 1 and Cluster 3. Cluster 2 highlights areas with a low male workforce proportion and very low female participation, which may contribute to the observed lower literacy rates. The histogram for Cluster-3 59 shows an extremely concentrated distribution of literacy percentage, with all values tightly clustered around a high percentage (around 80%). This suggests that states in this cluster have uniformly high literacy rates. Considering the demographics, Cluster 3 is an outlier and represents a state with an unusually high male workforce proportion and low female participation. This aligns with the homogeneous literacy trend observed.

The bubble plot 60 shows the relationship between male workforce participation and weighted female workforce participation across Indian states. The size of each bubble represents the weighting factor. States are spread across the plot, exhibiting a range of values for both male and female participation. Some key observations include the positive correlation between male and female participation, and outliers like Daman & Diu and Chandigarh with particularly high male participation compared to female participation.

The clustering in the image 61 reveals four distinct groups based on male workforce participation and weighted female workforce participation. Cluster 0 (red) represents states with high male participation and lower female participation, Cluster 1 (blue) contains states with lower participation for both genders, Cluster 2 (yellow) includes states with higher participation for both genders, and Cluster 3 (black) represents states with the highest participation for both genders. The clustering suggests a relationship between male and weighted female workforce participation.

e. **Feedback**

We realized that literacy is not the only factor influencing gender disparity in the workforce; urbanization also plays a significant role. This insight came from the visualizations we performed based on urbanization, which are presented in the next run. As a result, we decided to add columns related to urbanization to the dataset to better understand its impact on gender disparity in workforce participation.

3) The Third Run:

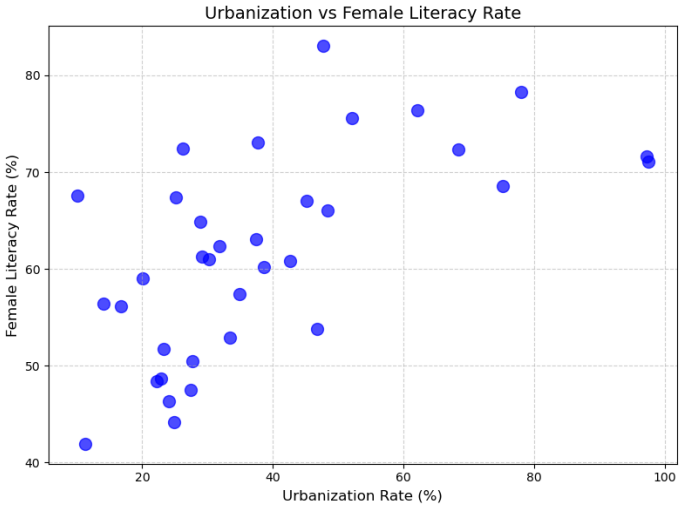

Fig. 62: Scatter plot depicting the relationship between urbanization and female literacy rate.

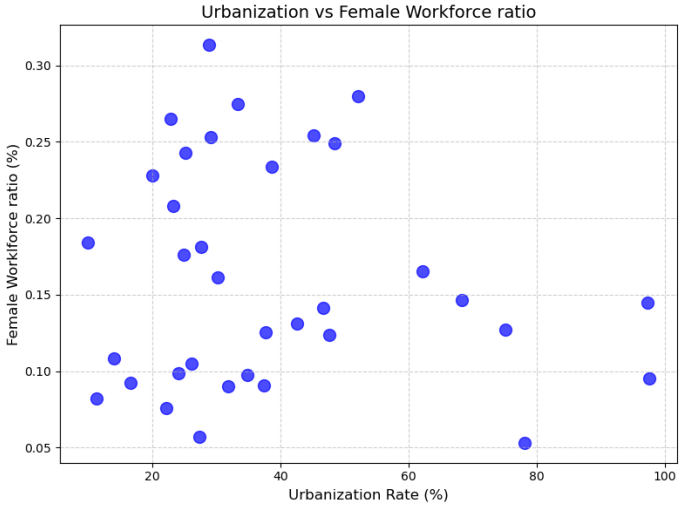

Fig. 63: Scatter plot illustrating the relationship between urbanization and female workforce participation ratio.

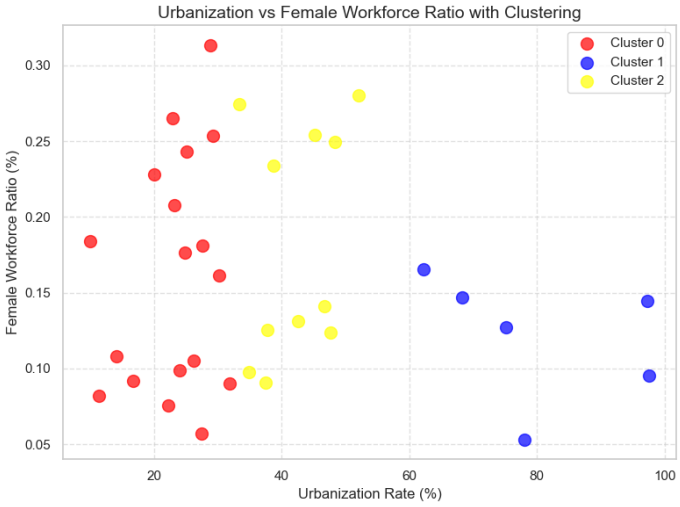

Fig. 64: Scatter plot of urbanization and female workforce participation ratio, with clusters identified based on these parameters.

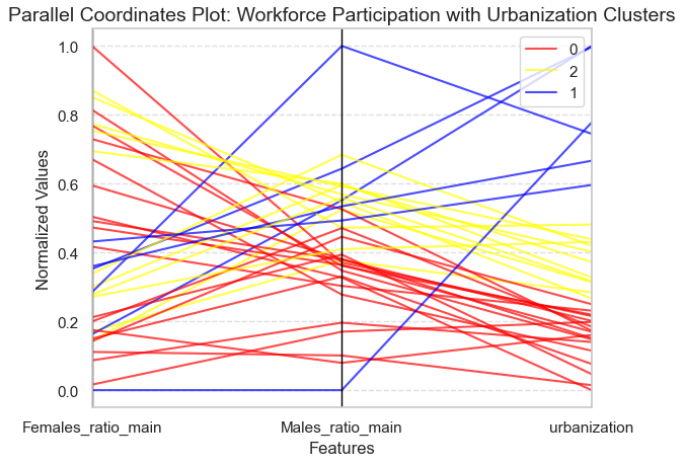

Fig. 65: Parallel coordinate plot visualizing male workforce ratio, female workforce ratio, and urbanization, with clusters represented by different colors.

a. **Data**
The dataset used for this analysis builds upon the previous iteration, with additional columns introduced to provide insights into urbanization. These new attributes include the *number of people residing in urban areas* and the *urbanization index*, which is calculated as the ratio of the number of people in urbanized regions to the total state population. Furthermore, the dataset retains the cluster assignments derived from the initial clustering analysis, facilitating a deeper exploration of patterns and relationships in workforce participation in conjunction with urbanization metrics.

b. **Models**
We again performed clustering using features such as female workforce participation and the urbanization index, as defined earlier. The objective of this analysis is to identify natural clusters based on urbanization and female workforce participation. Additionally, scatter plots before (63) and after (64) clustering are included in this report to visualize the differences and highlight the impact of clustering.

c. **Visualizations**
- Scatter plots (62) and (63) depict the distribution of urbanization with female literacy across states and the relationship between urbanization and female workforce participation. These plots help in understanding the influence of urbanization on both female literacy and workforce participation.
- A scatter plot (64) of urbanization versus female workforce participation after clustering is included

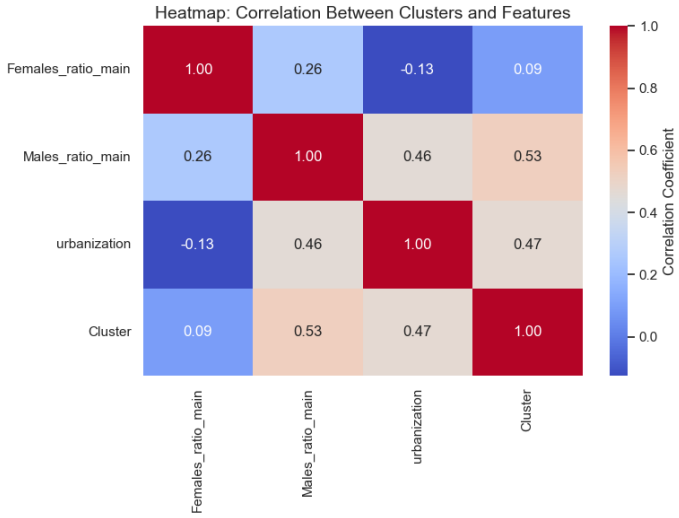

Fig. 66: Heatmap showing the correlation between female workforce participation ratio, male workforce participation ratio, urbanization, and clusters.

to identify natural groupings in the data.

- A parallel coordinates plot (65) illustrates the normalized values of female workforce participation, male workforce participation, and urbanization, offering insights into how these features interrelate.
- A heatmap (66) is provided to show the correlation between female workforce participation, male workforce participation, urbanization, and clusters, enabling a deeper understanding of their relationships.

d. **Knowledge**

The scatter plot 62 shows the relationship between the urbanization rate and the female literacy rate across different states in India. The plot reveals a positive correlation between these two variables, with states having higher urbanization rates generally exhibiting higher female literacy rates as well. The data points are spread out, indicating a wide range of urbanization and female literacy levels across the states. The trend line is upward sloping, confirming the positive correlation between urbanization and female literacy. States with higher urbanization (around 80-100%) have female literacy rates clustered between 70-80%. States with lower urbanization (around 40-60%) have a wider range of female literacy rates, from around 50% to 80%. This plot suggests that as states become more urbanized, there is a corresponding increase in female literacy rates, which could be an important factor in workforce participation and economic development.

The scatter plot 63 shows the relationship between urbanization rate and female workforce ratio, without the clustering information. There is a positive correlation between urbanization rate and female workforce ratio, with states having higher urbanization generally exhibiting higher female workforce ratios. The data points are spread out across the plot, indicating a wide range of values for both urbanization and female workforce ratio. The relationship appears to be linear, with no clear nonlinear or curvilinear patterns. The scatter of points suggests that the relationship between urbanization and female workforce ratio is not perfectly deterministic, as there is a fair amount of variation in the data. This plot provides a clear visualization of the underlying relationship between the two variables, without the additional information about the clustering of states. It helps to establish the general trend and range of values observed across the states.

The scatter plot 64 visualizes the relationship between urbanization rate and female workforce ratio, with the states color-coded by their assigned clusters. There is a general positive correlation between urbanization rate and female workforce ratio, with states having higher urbanization tending to have higher female workforce ratios. The clusters are well-separated, with Cluster 0 (red) having the lowest female workforce ratios, Cluster 1 (blue) having medium ratios, and Cluster 2 (yellow) having the highest ratios. Within each cluster, there is still a range of values for both urbanization rate and female workforce ratio, indicating diversity among the states in each cluster. The clustering seems to capture differences in the relationship between urbanization and female workforce participation, with the clusters representing distinct patterns or groups of states.

The parallel coordinates plot 65 provides a comprehensive view of the relationships between the three features (Female workforce Ratio, Male workforce Ratio, and urbanization) and the clustering of the states. There is a clear separation between the different clusters, with Cluster 0 (red) exhibiting the lowest values for both male and female workforce participation, as well as urbanization. Cluster 1 (blue) has higher values for all three features compared to Cluster 0, but lower values than Cluster 2 (yellow). Cluster 2 (yellow) represents the states with the highest values for male and female workforce participation and urbanization. The lines connecting the feature values for each state show the diversity within each cluster, with some overlap between the clusters, especially for the urbanization feature. This parallel coordinates plot offers a holistic understanding of how the workforce participation and urbanization characteristics are related and distributed across the different clusters of states. This visualization helps to understand how the clustering of states relates to the specific combination of urbanization and female workforce participation.

The heatmap 66 shows the correlation between the clusters and the three features: Female workforce Ratio, Male workforce Ratio, and urbanization. The diagonal elements represent the self-correlation of each feature, which is 1.0 as expected. The off-diagonal elements

show the correlations between the features. The Female workforce Ratio feature has a strong positive correlation (0.26) with the Male workforce Ratio feature, indicating a relationship between male and female workforce participation. The urbanization feature has a negative correlation (-0.13) with the Female workforce Ratio feature, suggesting that more urbanized states may have lower female workforce participation. The cluster feature has a positive correlation with the other three features, ranging from 0.09 to 0.53, indicating that the clustering is related to the workforce participation and urbanization characteristics. The heatmap provides a concise visualization of the relationships between the different features and the clustering of the states. This plot suggests that as states become more urbanized, there is a corresponding increase in female literacy rates, which could be an important factor in workforce participation and economic development.

e. **Feedback and Future Scope**
We could explore additional columns available in the dataset to identify potential correlations with workforce distribution. This can help uncover new insights and refine the current analysis, paving the way for more comprehensive studies in the future.

E. Summary

Our analysis focused on understanding the gender differences in workforce participation across Indian states. We used visualizations and clustering methods to learn more about these differences.

- **First run:** We analyzed male and female workforce participation across states. Visual tools like maps and scatter plots showed clear differences, and we grouped states into clusters. One key finding was that states with higher female workforce participation often had better literacy rates.
- **Second run:** We added literacy data to our analysis. This gave a deeper understanding of how literacy, especially among women, impacts workforce participation. The updated clusters helped us see patterns more clearly.
- **Third run:** We explored how urbanization affects workforce participation. Urbanized states generally had better female literacy and more balanced workforce participation.

Through this process, we gained important **knowledge** about how factors like literacy and urbanization influence workforce trends. The step-by-step approach showed that using the right data and tools can reveal patterns and insights that can help policymakers create better programs for education and gender equity.

V. MEMBER-WISE CONTRIBUTIONS

- **Aditya Priyadarshi:** He completely handled Task 1, "Analyzing Literacy-Driven Developmental Patterns", which involved implementing all aspects of the analysis. This included creating detailed visualizations to illustrate litteracy trends in India and applying clustering algorithms to identify and interpret patterns within the data. He also authored the corresponding report for Task 1, presenting the findings in a clear and structured manner.
- **Varnit Mittal:** He completely handled Task 2, "Analyzing Literacy and Socio-Economic-Dynamics in Female-Headed Households Through Census Data Visual Analytics", which involved implementing all aspects of the analysis. This included creating detailed visualizations to illustrate household trends and applying clustering algorithms to identify and interpret patterns within the data. He also authored the corresponding report for Task 2, presenting the findings in a clear and structured manner.
- **Ananthakrishna K:** He completely handled Task 3, "Analyzing Workforce Demographics in India", which involved implementing all aspects of the analysis. This included creating detailed visualizations to illustrate workforce trends and applying clustering algorithms to identify and interpret patterns within the data. He also authored the corresponding report for Task 3, presenting the findings in a clear and structured manner.

REFERENCES

[1] India Census dataset
[2] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Gorg,Jorn Kohlhammer, et al.. Visual Analytics: Definition, Process and Challenges.
[3] India Census Data by Government of India
[4] Columns in the unprocessed dataset
[5] Lloyd, S. P. (1957). Least squares quantization in PCM. **IEEE Transactions on Information Theory**, 28(2), 129-137.
[6] Andrew Ng, *How to Choose k in K-Means*, YouTube, 2024. Link to video
[7] Lin, C., Wu, J., & Zhang, H. (2013). Analyzing Cluster Transitions in Socio-Economic Data: Implications for Regional Development. Journal of Regional Analysis, 45(3), 124-138.
[8] Smith, R., Kumar, A., & Patel, S. (2017). Impact of Clustering Criteria on Socio-Economic Development Studies. International Journal of Economic Analysis, 52(1), 89-110.

VI. APPENDIX:

India Census 2011: Visual Insights

DAS 732: Data Visualization Assignment-1 Report

Varnit Mittal
IMT2022025
Varnit.Mittal@iiitb.ac.in

Aditya Priyadarshi
IMT2022075
Aditya.Priyadarshi@iiitb.ac.in

Ananthakrishna K
IMT2022086
Ananthakrishna.K@iiitb.ac.in

I. Introduction

This report analyzes the India Census 2011 dataset, focusing on three main areas. The dataset offers insights into various socio-economic and demographic factors, helping us identify patterns and relationships that shed light on the country's current situation. Our analysis is divided into the following tasks:

1) **Task 1:** Analysing Education Demographics and Their Interrelationship with various Factors
2) **Task 2:** Analysis of Housing Infrastructure and Planning in India
3) **Task 3:** Analysis of Workforce Demographics and Distribution in India

II. Preprocessing

The dataset contained numerous features, some of which were redundant. We selected the most relevant ones to derive meaningful insights. In 2011, Ladakh was not an independent Union Territory, and Telangana had not yet been established as a separate state. Since the data was provided at the district level, we grouped the districts that now belong to Telangana and updated their state name from Andhra Pradesh to Telangana. A similar update was made for the districts in Ladakh.

III. Tasks

A. *Task1: Analysing Education Demographics and Their Interrelationship with various Factors*

In this section, we focus on understanding how education levels across different states in India are influenced by factors such as household assets, internet access, workforce participation, and regional literacy rates. By visualizing these relationships, we aim to identify specific patterns, such as the impact of economic resources on higher education, the literacy divide between genders, and how non-main workers correlate with education dropouts. This analysis helps reveal how access to education varies across socio-economic and regional lines, providing insights that can guide targeted interventions to improve educational outcomes.

1) Literacy Rates : This section explores literacy rates across states, with a focus on the gap between male and female literacy levels. The visualizations highlight overall state-wise literacy patterns and compare female-to-male literacy ratios to overall rates

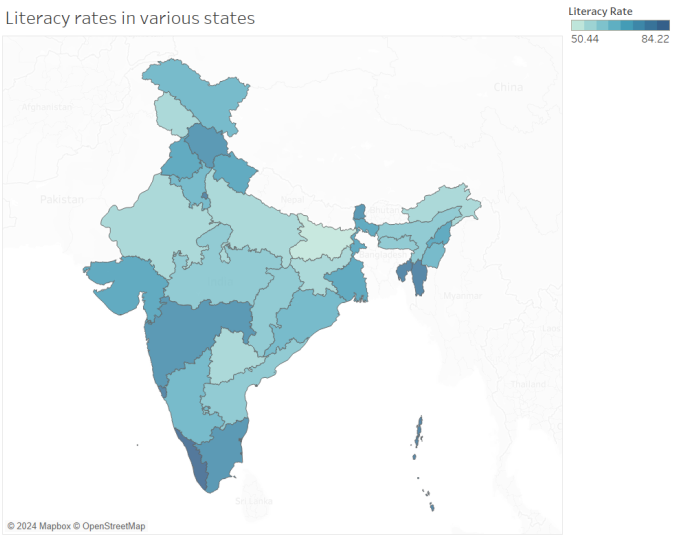

Fig. 1. Literacy Rates in Various States

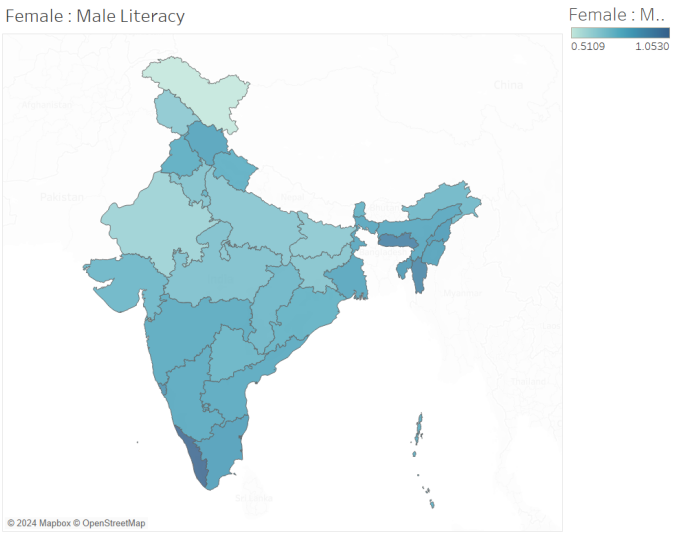

Fig. 2. Female : Male Literacy Rate

We created a choropleth map (Figure 1) to visualize literacy rates across various states of India. The map reveals that northern states such as Bihar, Jharkhand, and Uttar Pradesh; western states like Rajasthan; northeastern states like Arunachal Pradesh; and southern states like Telangana have lower literacy rates. In contrast, southwestern states like Maharashtra and Gujarat, southern states like Kerala and Tamil Nadu, northern states like Himachal Pradesh, Uttarakhand, and Punjab, northeastern states like Tripura, Mizoram, and Nagaland, and Union Territories like Delhi exhibit higher literacy rates.

While overall literacy rates provide a broad perspective, it is crucial to examine gender disparities in education. In Figure 2 and Figure 3, we present a map comparing male and female literacy rates across states. We observe that states such as Bihar, Uttar Pradesh, Jharkhand, Rajasthan, Haryana, and J&K have a low female-to-male literacy ratio ($\leq 70\%$), whereas states and UTs like Tamil Nadu, Goa, Pondicherry, Tripura, Mizoram, and Nagaland exhibit a higher ratio ($\geq 85\%$).

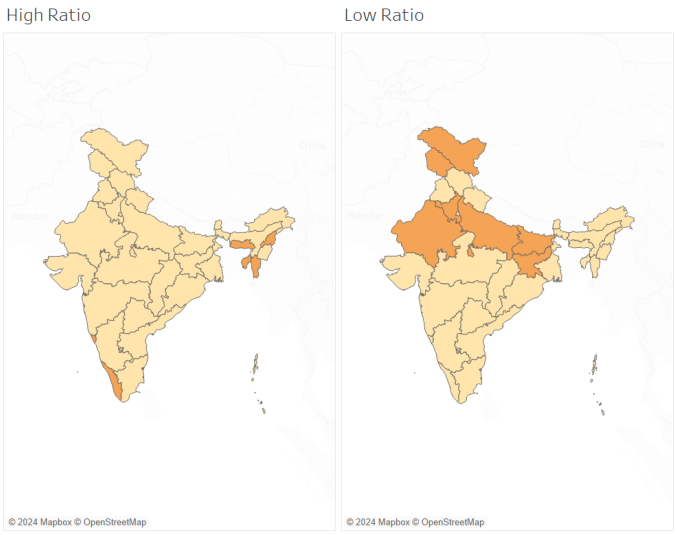

Fig. 3. Highlighting high and Low Female : Male Literacy Ratio states

The data shows that states with lower overall literacy rates tend to have lower female-to-male literacy ratios, and vice versa. This trend is further confirmed when plotting literacy rates against the female-to-male literacy ratio at the district level, revealing a strong positive correlation (Figure 4). This suggests that as overall literacy levels rise, gender equality in education also improves, leading to a more balanced literacy rate between males and females. Regions with higher literacy rates generally provide better educational opportunities for both genders, reducing the gender gap in literacy.

2) Level of Education: Building upon the analysis of literacy rates, it is equally important to explore the various levels of education attained across different regions.

In this subsection, we focus on the distribution of education levels, ranging from below primary to graduate education, and

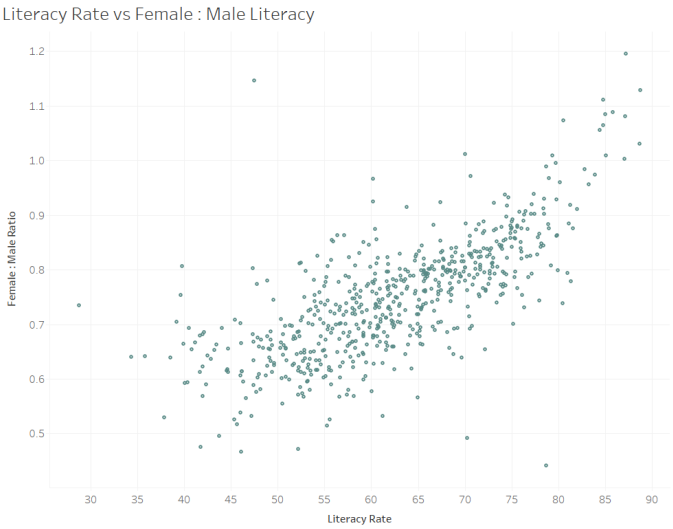

Fig. 4. Literacy Rate vs Female : Male Literacy Rate

analyze factors influencing progression and dropout at these stages. By examining this composition of education levels, we aim to uncover the relationship between dropout rates and workforce composition, as well as the impact of resources like internet access on higher education and beyond. This analysis sheds light on the socio-economic dynamics affecting educational attainment.

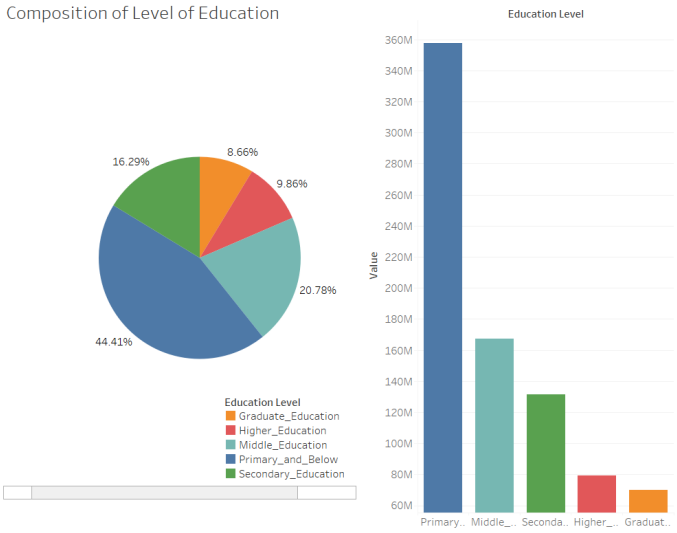

Fig. 5. Composition of Level of Education

The pie chart (Figure 5) reveals the percentage distribution of different education levels, with Primary and Below education forming the largest segment at 44.41%, followed by Middle Education at 20.78%, Secondary Education at 16.29%, Higher Education at 9.86%, and Graduate Education at 8.66%. The accompanying bar graph shows the absolute values for each level. A significant portion of students attend primary

school but do not continue to higher levels, dropping out at school `middle` school level. Similarly, many students discontinue their education after secondary school, leading to a smaller proportion pursuing higher education (*i.e.*, they dropout at `Higher` level). This pattern suggests a pyramidal structure, with widespread access to elementary education but significant barriers to advancing to higher levels. These drop-offs may be due to socio-economic factors, inadequate educational infrastructure, or / and cultural norms. Addressing these challenges is crucial to ensuring smoother progression through all levels of education and improving overall educational outcomes.

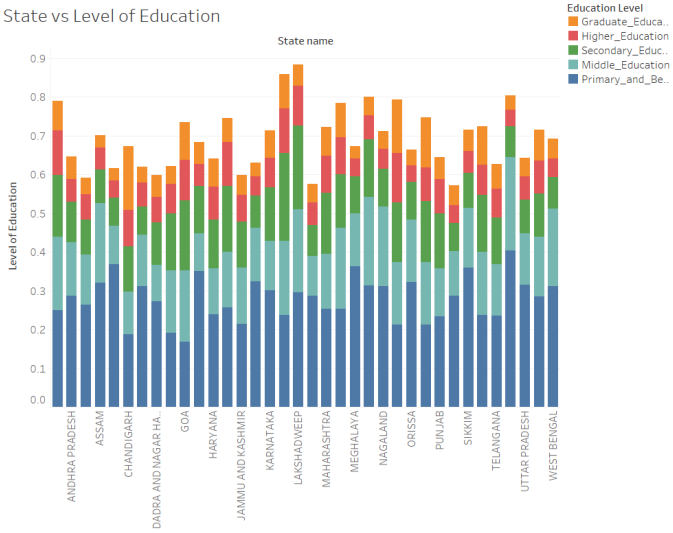

Fig. 6. State vs Level of Education

Building on this, the stacked bar graph (Figure 6) visualizes the distribution of education levels across different states. States like Goa, Maharashtra, and Punjab show a higher proportion of individuals attaining secondary and higher education, as indicated by the taller green and orange segments. In contrast, states such as Uttar Pradesh, West Bengal, and Orissa have a larger share of the population in the `Primary and Below` category, highlighted by the taller blue segments. This variation underscores disparities in educational attainment across states, likely reflecting differences in educational infrastructure, socio-economic conditions, or cultural norms.

Building on the analysis of educational attainment, we now explore the relationship between education levels and subsequent employment. From the Figure 7, we can see that states where students dropout at after their `Secondary` school or `High` school are Eastern and northeastern states, such as Bihar, West Bengal, Jharkhand, Assam, and Arunachal Pradesh, are depicted in darker shades. In contrast, Western and Southern states like Maharashtra, Karnataka, Andhra Pradesh, and Tamil Nadu show lower dropout rates. When examining the proportion of marginal and household workers, Eastern and northeastern states also have higher percentages, with states like Bihar, Jharkhand, Uttar Pradesh, Himachal

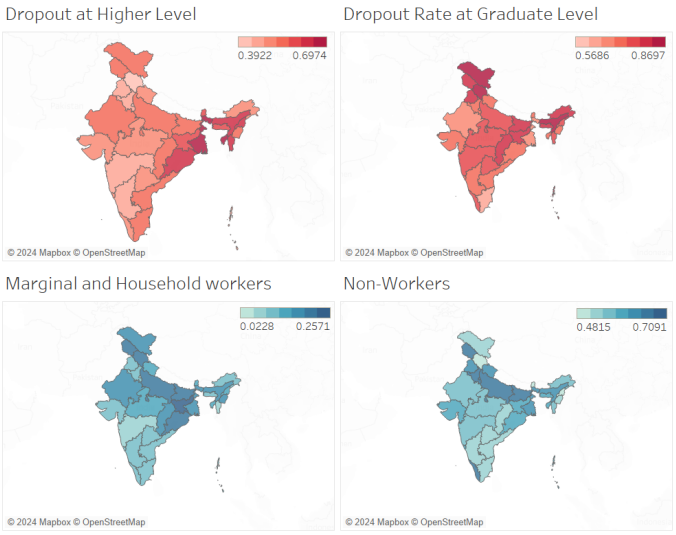

Fig. 7. Education Dropout against Worker composition

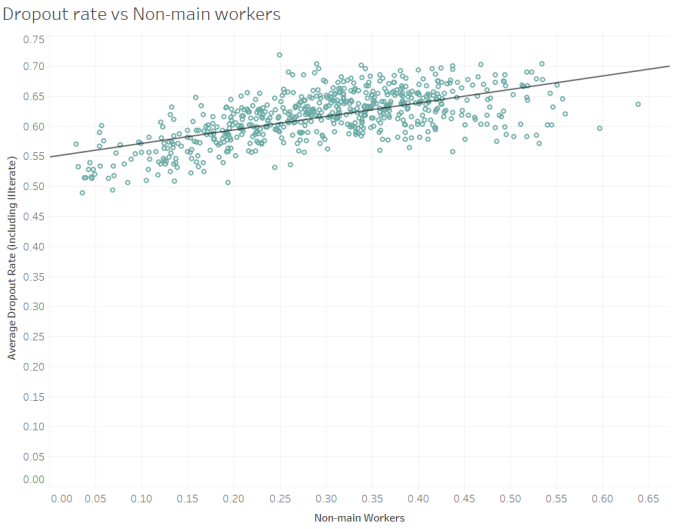

Fig. 8. Non main workers against Dropout Rate

Pradesh, and Odisha appearing in darker blues. This suggests that states with higher dropout rates tend to have higher proportions of individuals working as household workers, marginal workers, or being unemployed. Since state-level data provides a broader scope, we further analyzed the data at the district level to gain a more precise understanding. At this finer level of analysis, we observe that the ratio of non-main workers (i.e., household workers, marginal workers, and agricultural workers) increases as the dropout rate rises.

To further validate this observation, we plotted the number of main workers against the number of students pursuing higher education (Figure 9). This revealed a clear positive correlation, indicating that as the number of individuals opting for higher education increases, so does the number of main workers.

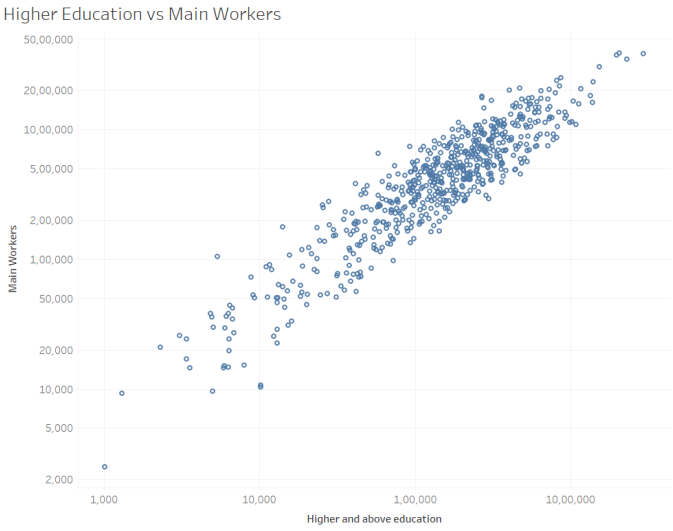

Fig. 9. Dependence of Higher Education on Number of `Main` workers.
Note: Plotted On Logarithmic Scale

This relationship can be explained by the fact that individuals with higher education are more likely to secure formal, stable employment, categorizing them as "main workers." Higher education provides the necessary skills and qualifications for more secure and higher-paying jobs, reducing reliance on marginal or household work. Conversely, lower levels of education often limit employment prospects, pushing individuals into informal sectors or leaving them unemployed. Thus, regions with lower dropout rates and better educational attainment tend to have a higher proportion of main workers, reinforcing the positive correlation observed in Figure 9.

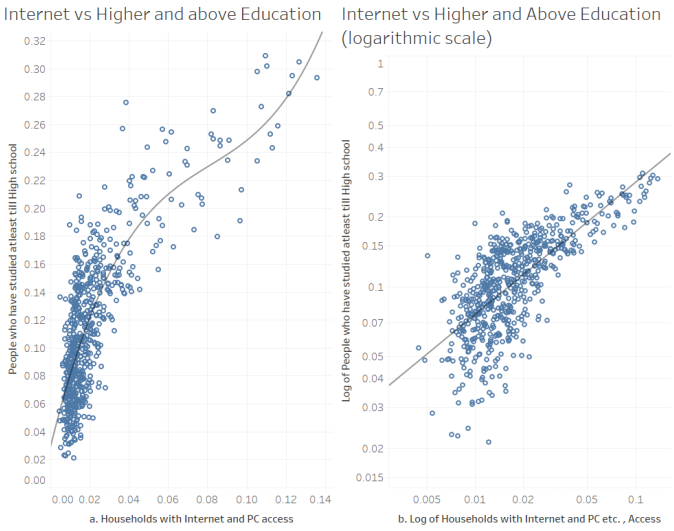

Fig. 10. Relation between number of student doing higher or above education and internet

We now examine the impact of internet access in households on higher education. Figure 10a illustrates that the number of students who have studied at least until high school increases more than linearly with internet access. To gain further insights, we analyzed the data on logarithmic scales and found a slope of 3 (Figure 10b). Since, $\log(y) \propto 3\log(x) \implies y \propto x^3$, this implies that an increase in the number of households with internet and computer access results in at least a tripling of the number of students who study up to at least high school. This relationship is explained by the fact that these technologies provide enhanced educational resources and opportunities which support student learning and academic persistence.

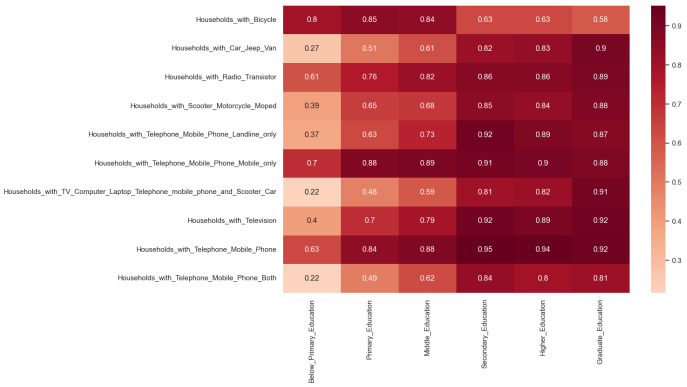

Fig. 11. Relation betweeen assest holding and education levels

3) Assets holdings: Continuing from our analysis of the relationship between internet access and educational outcomes, we now examine how educational attainment correlates with household asset ownership. The heatmap (figure 11) illustrates the correlation between household asset ownership and educational attainment. The assets represented include bicycles, cars, radios, mobile phones, televisions, and more. Households with lower education levels, such as below primary or primary education, tend to possess more basic assets, like bicycles, radios, and a single mobile phone. In contrast, as education levels rise—from secondary to graduate education—the ownership of more advanced assets, such as cars, televisions, and multiple mobile phones, becomes more prevalent.

Notably, higher education levels strongly correlate with modern amenities, including televisions and both mobile and landline communication devices. This heatmap effectively illustrates the link between higher educational attainment and household wealth, highlighting how education influences access to more valuable and modern assets, thereby reflecting its impact on socioeconomic status and living standards.

*B. **Task 2**: Analysis of Housing Infrastructure and Planning in India*

This section provides a comprehensive analysis of housing infrastructure and planning in India, focusing on the current state of urban and rural housing, and challenges related to land use, affordability, and sustainable development. By visualizing various parameters, we offer to seek insight into housing initiatives are aligned with population needs and

future urban planning strategies. We aim to visualize spatial and demographic patterns in housing development and assess the impact of urbanization.

1) Density of Owned and Rented Houses : We created two choropleth maps (fig. 12) to visualize the density of owned and rented houses across various states of India, offering valuable insights into the distribution of housing tenure types. The map on the left illustrates regions with high densities of home-ownership, highlighting states where owning property is the predominant form of housing. In contrast, the map on the right displays areas where renting is more prevalent, particularly in regions with more urbanized populations and greater housing demand due to migration and job opportunities.

This spatial representation allows for a deeper understanding of how housing preferences vary by region, reflecting economic, cultural, and demographic factors that shape the housing market in India.

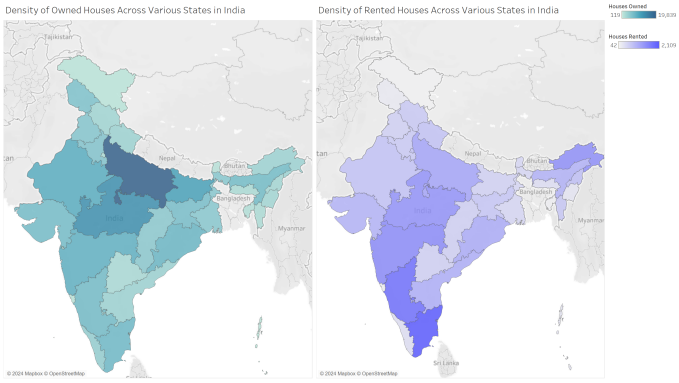

Fig. 12. Density of Owned and Rented Houses.

States with higher densities of owned houses, such as Uttar Pradesh and Bihar, are often characterized by rural populations and a strong cultural inclination towards property ownership. On the other hand, states like Tamil Nadu and Karnataka, which exhibit a greater density of rented housing, are more urbanized, with vibrant economies attracting a significant portion of the population seeking rental accommodations.

From a policy perspective, the visualization can inform housing strategies and urban planning. For instance, **National Housing Policy (1988)** and **Rent Control Act** have aimed to bridge the gap between housing supply and demand, focusing on affordable housing. The visualization of ownership and rental densities offers a critical lens through which to assess whether these initiatives are meeting the population's needs in different regions.

The grouped-bar chart analysis (fig. 13) further underscores the need for data-driven housing policies that account for regional variations in housing preferences, helping align future housing and urban planning strategies with the dynamic and evolving needs of India's growing population.

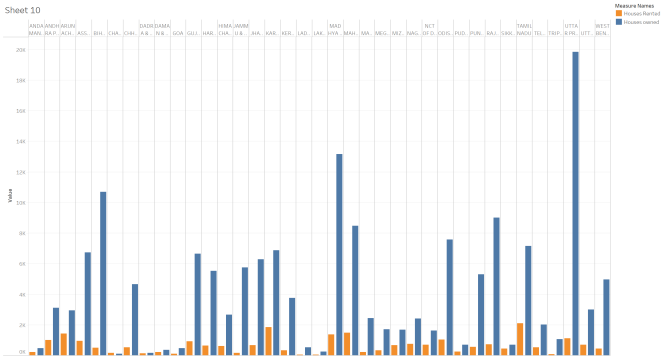

Fig. 13. Statewise Owned and Rented Houses.

2) Household Sizes : We have created a stacked bar chart (fig. 14) which aims to visualize the distribution of household sizes across various states in India, offering critical insights into family structure and demographic trends. The chart breaks down households by size, ranging from small families with two members to larger families with nine or more members. This data serves as an essential indicator for understanding regional differences in family dynamics and their implications for housing and resource allocation.

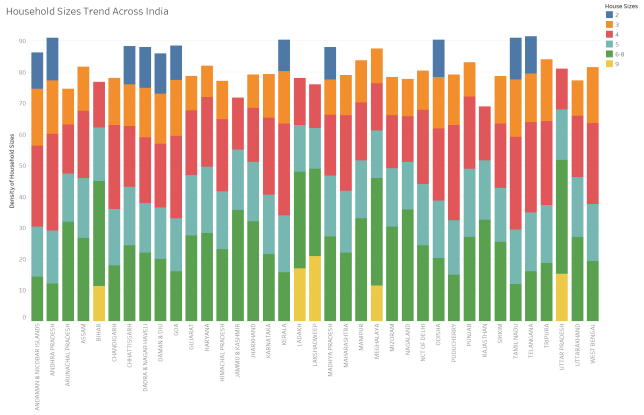

Fig. 14. Household Sizes Trend across India.

Larger household sizes, such as those seen in states like Bihar and Uttar Pradesh, tend to reflect cultural norms around joint families, where multiple generations live together. Conversely, states with higher densities of smaller households, such as Goa and Kerala, indicate a trend towards nuclear families, which is often associated with urbanization, higher literacy rates, and economic independence.

From a policy perspective, understanding these variations is essential for designing effective housing initiatives and social programs. Larger households require more living space, creating a demand for housing units that can accommodate extended families. In contrast, regions with smaller household sizes may focus more on affordable housing options for nuclear families, which tend to prefer compact and efficient housing solutions in urban environments.

We have also created a grouped-bar chart (fig. 15) to delve further into this. This visualization clear shows the trend of large v/s small household sizes across various states in India in a decreasing fashion.

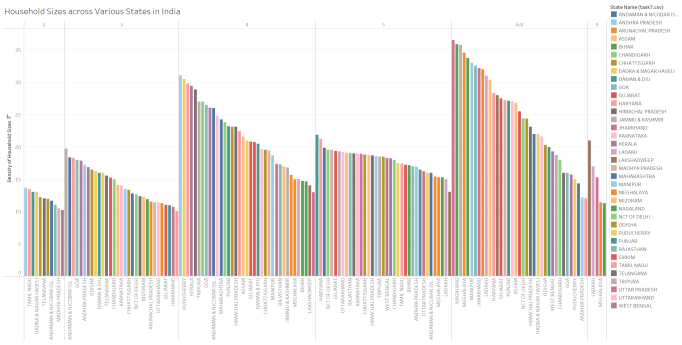

Fig. 15. Statewise Trend in Household Sizes in a Decreasing Fashion.

The alignment of household sizes with housing development plays a critical role in promoting sustainable growth. Regions with a predominance of larger families face challenges related to housing shortages, overcrowding, and the strain on local infrastructure. Meanwhile, areas with smaller households may see increasing demand for multi-family units or apartments, which could lead to higher densities in urban centers. Policymakers must ensure that housing projects are flexible enough to cater to both extremes, balancing land use with sustainable growth.

By aligning urban planning and infrastructure development with the observed household size distribution, government initiatives can more effectively address regional needs while supporting the broader goals of sustainability.

3) Drinking Water Source Distances : We have created a pie-chart (fig. 16(a)) and a packed bubble diagram (fig. 16(b)) to visualize this data across India. The provided charts offer a comprehensive visualization of the distribution of water source distances from households across India, giving essential insights into water accessibility. The data is categorized into three distinct groups: households with water sources "Within Premises," those "Near Premises," and those located "Away" from their homes. This information is crucial for understanding the regional disparities in access to water and its broader implications for public health, infrastructure, and socio-economic conditions.

Households with water sources within premises represent the most ideal scenario, as these families have direct access to drinking water. This category reflects the success of infrastructural developments, particularly in urban and semi-urban areas, where piped water connections are more prevalent. As seen in states with higher urbanization rates, access to water within the premises is not only a convenience but also a significant factor in reducing health risks associated with waterborne diseases and ensuring better hygiene standards.

On the other hand, households with water sources near premises require a moderate amount of effort to access water.

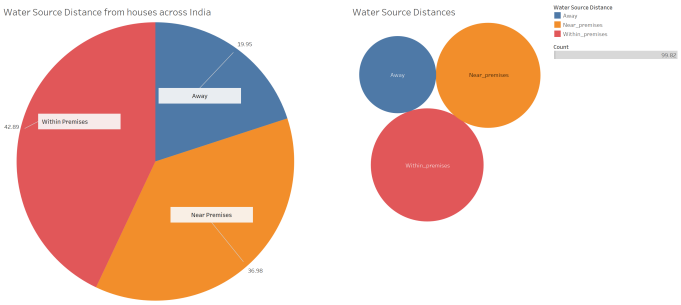

Fig. 16. Drinking Water Source Distance.

This group highlights a common situation in rural or peri-urban areas, where centralized water systems may be present, but the distribution network does not yet extend directly into homes. The burden of water collection in these households still exists but is less severe, particularly in terms of the time and labor required.

The most challenging situation is faced by households with water sources categorized as away from premises, which are predominantly located in rural areas with underdeveloped water infrastructure. In these cases, individuals, often women and children, must travel considerable distances to fetch water, a task that significantly impacts their time, energy, and quality of life. Moreover, the long distance to water sources increases the risks of contamination during transportation and the likelihood of relying on unsafe water sources, leading to adverse health outcomes.

The following Sun-burst diagram (fig. 17) will give you a better prospective of State-wise distribution of drinking water supplies. This highlights the fact that Urbanization is caused due to lower rates of development in rural areas or states where the rural areas are more dominant.

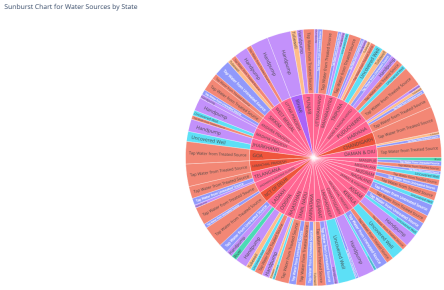

Fig. 17. Statewise Drinking Water Source.

This chart visualizes the distribution of water sources across various states in India, breaking down water sources into categories such as tap water from treated and untreated sources, handpumps, uncovered wells, and others. Each segment represents the proportion of households in a particular state relying on a specific water source. States like Goa, Himachal Pradesh, and others show a higher reliance on treated tap water, reflecting better infrastructure and water

management systems. On the other hand, states like Uttar Pradesh and Bihar show significant reliance on handpumps and untreated sources, highlighting gaps in water quality and accessibility. This chart offers an essential view of the disparities in water access across regions, which can have broader implications for public health and socio-economic development.

4) Latrine Availability Across Indian States : We made a scatter plot (fig. 18) which aims to visualize the distribution of households with and without latrines inside their premises across various states in India, providing critical insights into sanitation infrastructure. The markers differentiate between households that have latrines within their premises (squares) and those without latrines inside premises (circles).

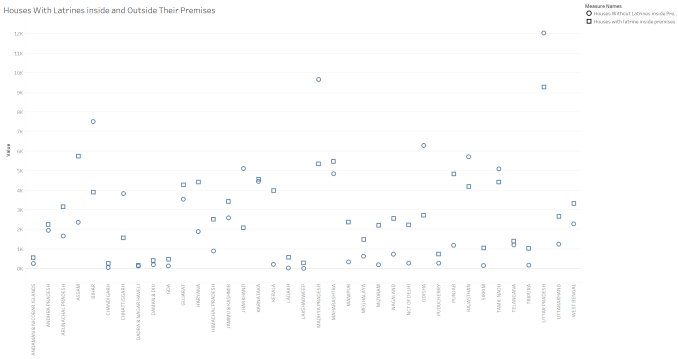

Fig. 18. Statewise Latrine Distribution.

From a national perspective, the data underscores the regional disparities in sanitation access, which is a key determinant of public health and well-being. States such as Maharashtra and Bihar exhibit significant numbers of households without latrines, which reflects ongoing challenges in improving sanitation coverage. In contrast, Kerala and Himachal Pradesh demonstrate a higher proportion of households with latrines inside their premises, indicating better sanitation infrastructure, often attributed to more effective policy interventions and public health campaigns.

The scatter plot further highlights that states with higher number of rural and remote regions face greater challenges, with more households lacking private sanitation facilities. The absence of latrines inside premises not only contributes to poor health outcomes, especially in vulnerable populations, but also perpetuates social inequalities.

By addressing the challenges of latrine availability, particularly in regions that lag behind in sanitation coverage, policymakers can work towards improved public health outcomes and a more equitable distribution of resources.

5) Rural and Urban Living Conditions : We have created a stacked bar-chart (fig. 19) which aims to present a comprehensive view of the disparities in livable housing conditions across various Indian states and union territories, with a focus on rural and urban areas. Each state is represented by a pair of stacked bars, where the blue section represents rural living conditions, and the orange section reflects urban living conditions. This visualization offers critical insights into the statewise divide in housing infrastructure and accessibility.

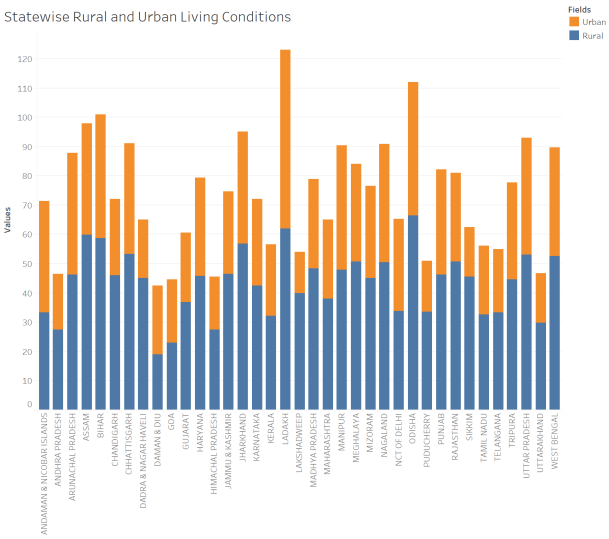

Fig. 19. Statewise Distribution of Housing Conditions in Rural and Urban Areas.

One of the key takeaways from this chart is the pronounced urban-rural divide. Across most states, urban areas demonstrate better livable conditions, as indicated by the higher values in the orange sections. For example, Maharashtra and Gujarat display a substantial urban advantage in housing quality, while states like Bihar and Uttar Pradesh have relatively lower values across both rural and urban areas. This trend suggests that certain states face more widespread challenges, particularly in improving housing infrastructure, even in urban settings.

At the same time, the chart reveals notable state-specific variations. In states like Kerala and Himachal Pradesh, the gap between rural and urban areas is relatively narrow, suggesting more balanced housing policies that cater to both demographic groups. On the other hand, states such as Bihar and Jharkhand show a stark contrast between rural and urban areas, reflecting deep-rooted infrastructure disparities that may be driven by uneven distribution of resources and underdeveloped rural regions.

In highly urbanized regions like Delhi, Puducherry, and Maharashtra, the concentration of high urban values reflects the focus on economic development in urban centers. These states have benefited from rapid urbanization and industrial growth, which has led to better infrastructure in cities. However, this pattern also underscores the neglect of rural areas, where livable conditions remain a significant challenge.

In conclusion, the chart highlights the substantial disparities in livable housing conditions between urban and rural India. From a policy perspective, this chart provides a roadmap for targeted interventions. States like Uttar Pradesh, Bihar

and Jharkand require investments in housing, sanitation, and basic utilities to bridge the gap with their urban counterparts. Moreover, regions with high urbanization may need policies that ensure the rural population is not left behind during rapid economic development.

6) Material Used to Build Walls of Houses across India : We have used a treemap visualization (fig. 20) to illustrate the distribution of wall-building materials used across India, providing a comprehensive view of the diverse materials employed in housing construction. The chart highlights a significant reliance on locally sourced, traditional materials, as well as modern alternatives that reflect both economic and geographic factors.

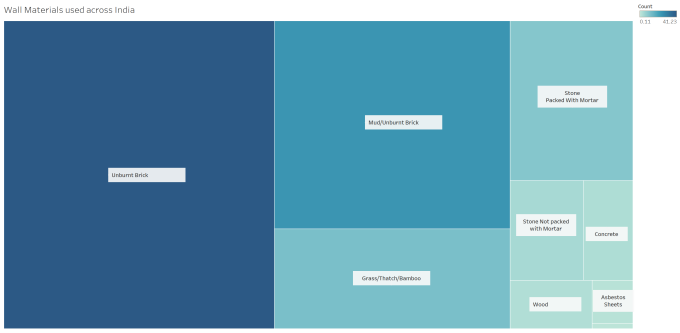

Fig. 20. Wall-Building Materials Used Across India.

The most dominant category in this treemap is "Unburnt Brick," which occupies a substantial portion of the visual space, indicating that this material is widely used across the country. This is likely due to its cost-effectiveness, ease of availability, and suitability for various climates. The second-largest category, "Mud/Unburnt Brick," represents an even more basic construction material, typically used in rural areas where industrial materials are either expensive or inaccessible. This suggests that a large segment of India's population still relies on traditional building techniques, particularly in less urbanized regions.

Other materials such as "Stone Packed with Mortar" and "Stone Not Packed with Mortar" reflect the influence of geographical diversity, where stone is abundant and labor costs for packing materials are variable. "Concrete" also appears in the chart, though in smaller quantities, indicating that while modern construction methods are gaining ground, they are still secondary to traditional materials in many parts of the country.

We have also used packed bubble diagram (fig. 21) which provides and insightful visualization of the diverse materials used in house construction across various states and union territories of India. Each bubble represents the proportion of households using a specific material for their walls, with the size of the bubble indicating the relative prevalence of that material within a state. Different colors in the diagram correspond to different materials such as burnt brick, concrete, stone, bamboo, mud, and wood, offering a clear snapshot of regional variations in construction practices.

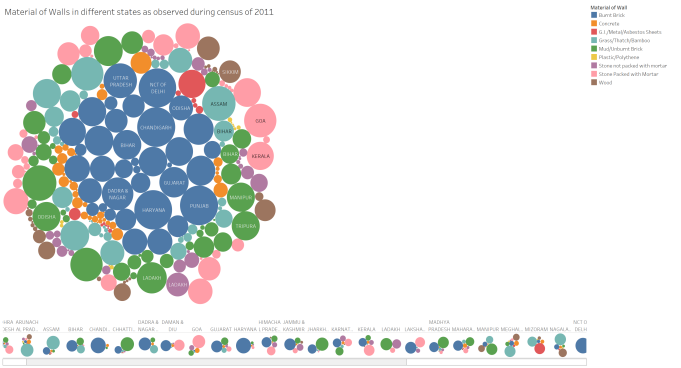

Fig. 21. Statewise Distribution of Wall-Building Materials.

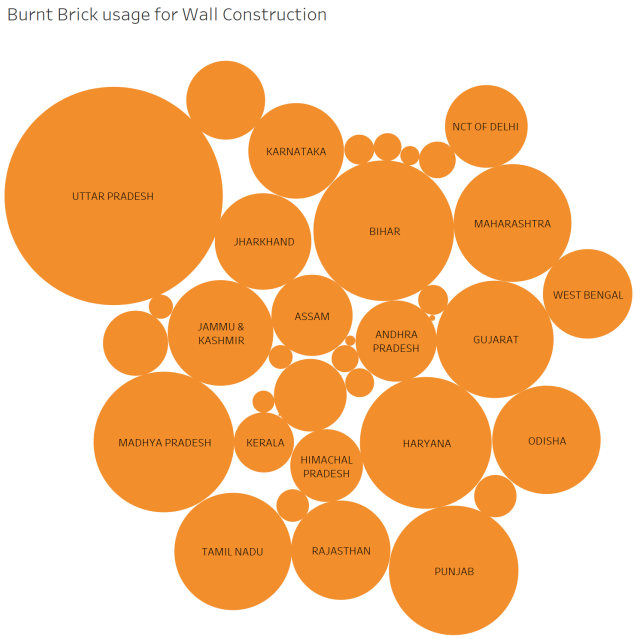

Fig. 22. Statewise Distribution of Usage of Burnt Brick.

One of the most notable aspects of the diagram is the regional dominance of certain materials. For instance, states like Uttar Pradesh, Bihar, and Gujarat are characterized by large blue bubbles, reflecting the widespread use of burnt bricks as the primary construction material for walls. This suggests that bricks are a dominant choice in these states due to their availability and durability. This can be seen in (fig. 22). The dominance of burnt brick across India as observed during **Census of 2011** signifies the fact that despite the availability of modern construction material like concrete, people still prefer locally available and cheap materials like burnt brick, bamboo etc. This can also be observed from fig. 21 where north-eastern states like Assam and Manipur,

show a higher reliance on bamboo and thatch, that are easily sourced from local environment.

7) Material Used to Build Roofs of Houses across India : We have created a treemap (fig. 23) to provide a visual representation of the different types of roofing materials used in residential structures throughout the country. Each section of the treemap corresponds to a specific material, with the size of each rectangle indicating the relative proportion of houses utilizing that material. The color gradient represents the count of households that use a particular material, with darker shades indicating a higher prevalence.

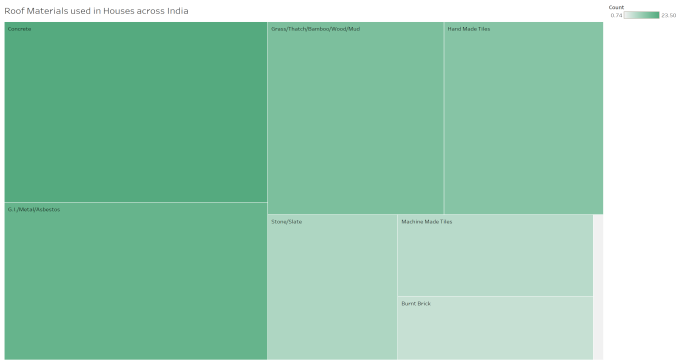

Fig. 23. Roof-Building Materials Used Across India.

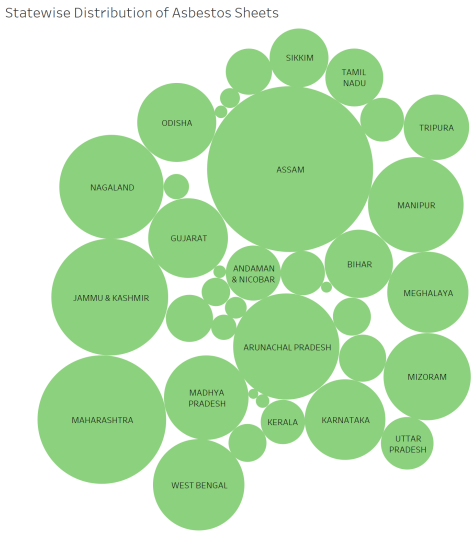

Fig. 24. Statewise Distribution of Usage of Asbestos Sheets.

Concrete emerges as the most dominant roofing material in the dataset, occupying the largest section of the treemap. This is reflective of the increasing trend toward modern, durable construction methods in urban and semi-urban areas, where concrete is often preferred due to its strength, longevity, and resistance to weather-related wear and tear. Concrete roofs are also fire-resistant and provide better thermal insulation, making them a desirable choice in many parts of India with varying climates.

Another prominent material displayed is G.I./Metal/Asbestos as seen in (fig. 24). This material is commonly used in both urban and rural areas, particularly for its affordability and ease of installation. However, asbestos usage raises public health concerns due to its known carcinogenic properties, underscoring the need for regulatory measures to phase out asbestos in favor of safer alternatives.

A large proportion of households still use Grass/Thatch/Bamboo/Wood/Mud as roofing materials, especially in rural and remote areas. These materials are often locally sourced, making them cost-effective for lower-income households. However, their usage points to concerns regarding durability and vulnerability to weather conditions, particularly during the monsoon season. Roofs made from natural materials are also susceptible to quicker degradation over time and provide less insulation compared to more modern alternatives.

Understanding the distribution and choice of roofing materials is crucial for the development of housing and infrastructure policies. By promoting sustainable and safe building practices—such as replacing asbestos with safer materials, improving access to affordable and durable roofing options in rural areas, and ensuring that urban development is environmentally friendly.

8) Cooking Fuel Usages in Rural and Urban India : We have created a circle-view diagram (fig. 25) which illustrates the varied use of cooking fuels in rural and urban households across India. The chart categorizes fuel usage by type—such as LPG/PNG, firewood, crop residue, cow dung cake, kerosene, and charcoal—and visualizes their prevalence through the height of the circles, with higher circles indicating higher usage.

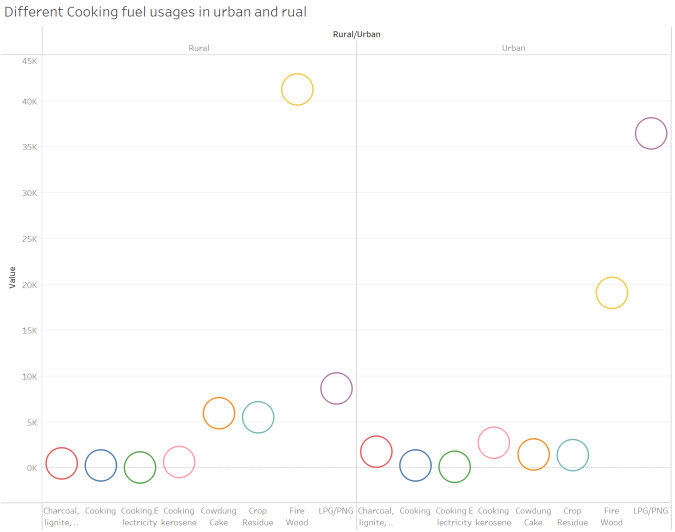

Fig. 25. Cooking fuel usages in urban and rural India.

In rural areas, the data highlights a significant reliance on traditional fuels such as firewood, cow dung cakes, and crop residue, with firewood being the dominant fuel source. These traditional fuels, though widely available and cost-effective, have serious environmental and health implications, contributing to deforestation and indoor air pollution, which is a leading cause of respiratory diseases in rural households. The circles with less values representing LPG/PNG in rural areas indicate its relatively limited adoption.

In contrast, urban households predominantly use LPG/PNG, as circles with higher values in this category. This shift toward cleaner fuel sources in urban areas can be attributed to better infrastructure, higher incomes, and greater access to government-subsidized LPG connections. The usage of firewood and other traditional fuels in urban settings is minimal, indicating a move toward modern energy sources that are safer and more efficient.

The visual comparison between rural and urban areas underscores a critical energy access gap. While urban households are rapidly adopting cleaner cooking fuels, rural households remain reliant on polluting fuels, highlighting the need for continued policy efforts to improve access to affordable and clean cooking energy in rural regions. This gap also points to underlying socio-economic disparities, as rural communities often lack the resources and infrastructure needed to transition to cleaner energy options.

*C. **Task 3**: Analysis of Workforce Demographics and Distribution in India*

This section provides a comprehensive analysis of workforce demographics and distribution in India, with a focus on understanding gender composition, employment categories, and regional patterns of workforce participation.

We created two choropleth maps to visualize the percentage of females and males in the workforce across Indian states (Figures 26 and 27). These maps provide a comprehensive overview of workforce participation patterns throughout the country. Northeastern states like Manipur and Nagaland, as well as northern states such as Rajasthan and Madhya Pradesh, show higher percentages of female workforce participation. Many northern and western states exhibit relatively lower percentages of female workforce participation. The higher female workforce participation in northeastern states could be attributed to their predominantly agricultural economies. Women in these regions often engage in agriculture and allied activities, contributing to increased female workforce participation.

To further analyze the data, we created a bar chart illustrating female workforce participation ratios across Indian states and union territories (Figure 28). Northeastern states (Nagaland, Manipur, and Meghalaya) and states like Himachal Pradesh and Chhattisgarh show significantly high gender ratios. Union territories such as Daman and Diu and Lakshadweep have low gender ratios, possibly due to their smaller populations. Major regions with low gender ratios include Delhi, Punjab, Chandigarh, and West Bengal.

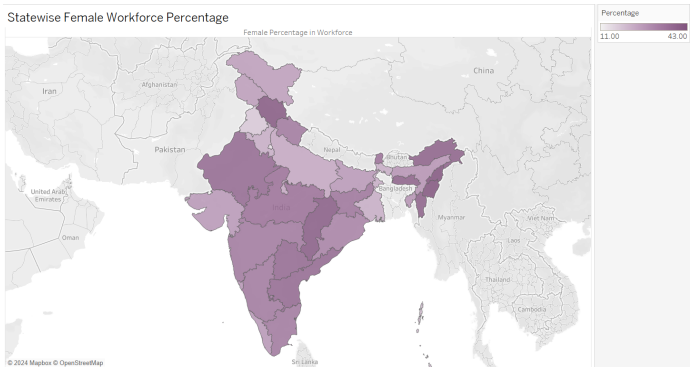

Fig. 26. Statewise Female Worker Percentage.

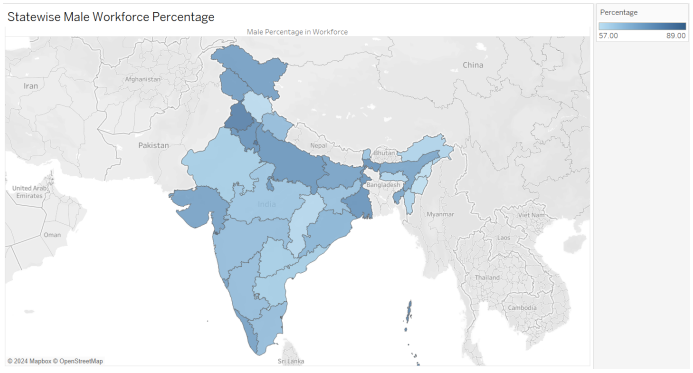

Fig. 27. Statewise Male Worker Percentage.

Having examined the gender composition of the workforce across different Indian states, we now turn our attention to the overall structure of the Indian workforce, regardless of gender. This broader perspective allows us to understand how the total working population is distributed across different types of employment engagement. To visualize this, we created a pie chart(Fig 29) that categorizes workers into three main groups: main workers, marginal workers, and non-workers.

The pie chart reveals a striking distribution: non-workers form the largest segment, comprising approximately 50-55% of the eligible workforce. Main workers, those in full-time employment, make up the second-largest group at around

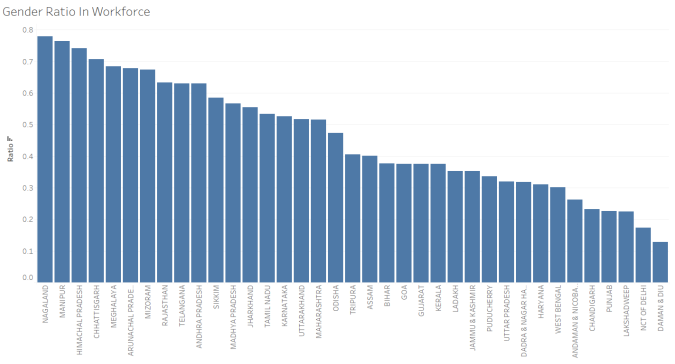

Fig. 28. Statewise Gender Ratio.

30-35%. Marginal workers, who are employed part-time or seasonally, represent the smallest category at roughly 10-15%. The high proportion of non-workers suggests significant economic challenges, possibly including high unemployment or underemployment rates. The substantial segment of marginal workers indicates the presence of an informal or gig economy.

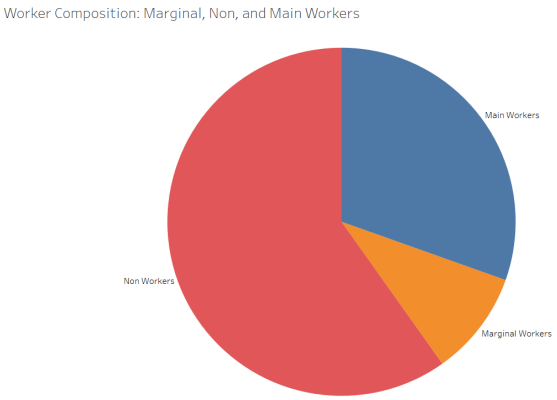

Fig. 29. Worker Composition: Main, Marginal, Non.

To delve deeper into this composition, we created three separate bubble charts (Fig 30, 31, 32) to illustrate the statewise distribution of each worker category.

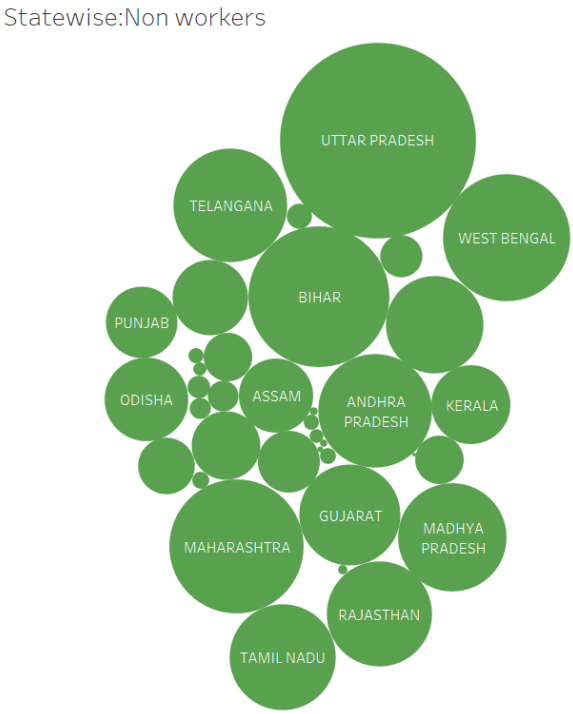

Fig. 30. Statewise Worker Composition: Non Workers

The bubble chart for non-workers highlights significant regional variations. Uttar Pradesh stands out with the largest bubble, indicating it has the highest number of non-workers among all states. This is followed by Bihar and West Bengal. States like Maharashtra, Madhya Pradesh, and Rajasthan also show substantial non-worker populations. This pattern may be attributed to factors such as high unemployment rates, large student populations, or cultural norms that limit workforce participation in these regions.

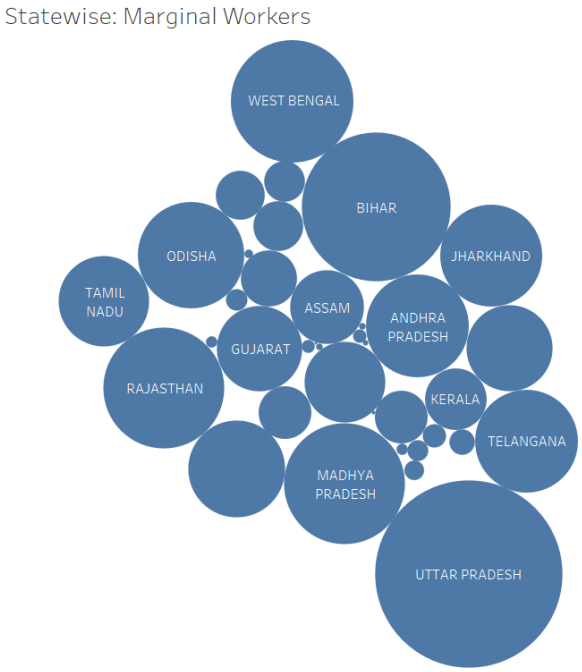

Fig. 31. Statewise Worker Composition: Non Workers

The marginal workers chart reveals a slightly different pattern. While Uttar Pradesh still shows the largest bubble, Bihar follows closely. States like West Bengal, Madhya Pradesh, and Jharkhand also have notable marginal worker populations. This distribution could be indicative of seasonal employment patterns, particularly in agriculture-dependent regions, or a prevalence of informal and part-time work arrangements in these states.

The main workers chart presents a more diverse picture. Although Uttar Pradesh maintains a large bubble due to its population size, states like Maharashtra, Andhra Pradesh, Telangana, Karnataka and Tamil Nadu show comparatively larger bubbles than in the previous charts. This suggests that these states may have more developed formal employment sectors or industries that provide full-time employment opportunities.

In addition to analyzing the overall composition of the Indian workforce in terms of main, marginal, and non-workers, we also examined the workforce divided by the type of employment: Agricultural Workers, who are employed as laborers on farms or agricultural lands; Cultivator Workers, who own or lease land and cultivate crops; Household Workers, typically involved in domestic duties within households; and Other Workers, covering all other forms of employment, such as industrial and service sectors. The pie chart (Fig 33) presents the distribution of the workforce across these four categories,

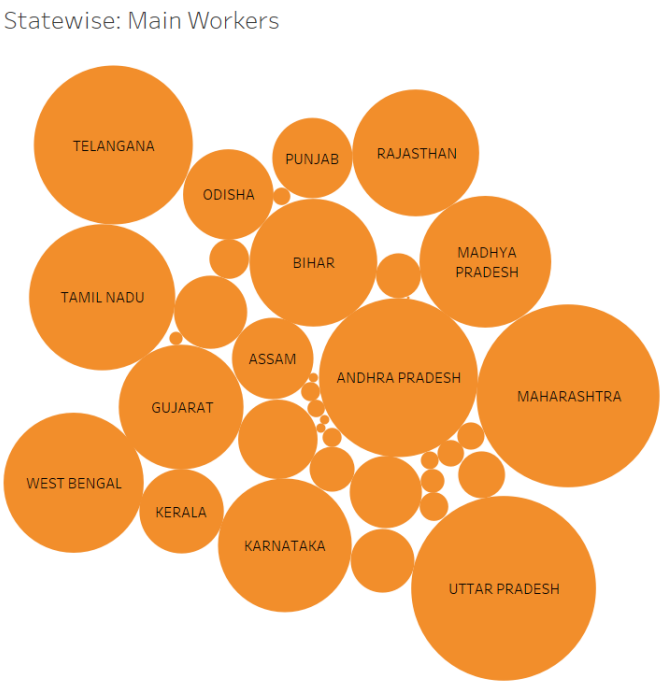

Fig. 32. Statewise Worker Composition: Non Workers

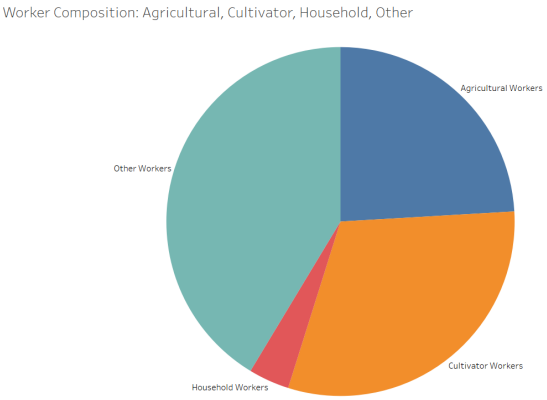

Fig. 33. Worker Category: Agricultaral, Cultivator, Household

providing a snapshot of the employment landscape. The largest segment is Other Workers, suggesting that a significant portion of the population is engaged in sectors outside of agriculture or household work. This could include a mix of industrial, service-oriented, and informal sector jobs, indicating the shifting nature of employment in India, particularly as urbanization and industrialization continue to expand. Agricultural Workers and Cultivator Workers together form a notable portion of the workforce, reinforcing the role of agriculture as a key economic driver. Household Workers make up the smallest category, likely representing domestic labor and care giving roles, which tend to be informal and under-reported sectors in the workforce data.

To further explore these categories on a state-by-state basis, a stacked bar chart(Fig 34) was created to visualize the workforce composition across different regions. The chart reveals significant regional variation, with some states exhibiting a strong reliance on agriculture while others show a more diversified workforce.

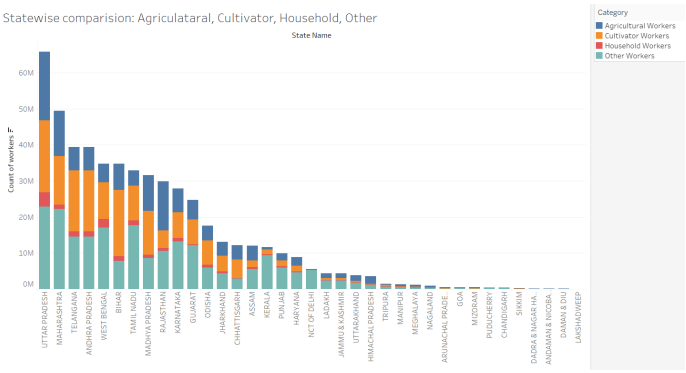

Fig. 34. Statewise Worker Category: Agricultaral, Cultivator, Household

Uttar Pradesh has the largest total workforce, reflected by its significant number of both agricultural and other workers. States like Maharashtra and Tamil Nadu show a much larger proportion of Other Workers, indicating a shift towards non-agricultural employment sectors. These states are more urbanized and industrialized, which could explain their greater concentration of non-agricultural jobs. In contrast, states like Telangana and Andhra Pradesh exhibit large segments of Cultivator Workers, which highlights the ongoing importance of farming and agriculture in these regions.

States like Maharashtra and Tamil Nadu, with a higher proportion of Other Workers, may be experiencing greater industrial growth and urban employment opportunities, while states like Uttar Pradesh and Bihar continue to rely heavily on agriculture as a major source of employment. These patterns reflect broader socio-economic trends and regional disparities within India's economy.

Continuing from the analysis of the worker categories, we now take a closer look at the relationship between Cultivator Workers and Agricultural Workers through a scatter plot(Fig 35), which illustrates the state-wise comparison between these two employment sectors.

Uttar Pradesh stands out with the largest numbers in both categories, showcasing its strong dependence on agriculture. In contrast, Bihar and Andhra Pradesh have a higher number of Cultivators, indicating a landowner-driven agricultural economy, where more individuals work their own land. This contrasts with Rajasthan, where the workforce is more skewed toward Agricultural Workers (laborers), suggesting a larger labor-driven agricultural sector.

Maharashtra and Madhya Pradesh also reflects a predominance of agricultural laborers, reinforcing the state's reliance on hired farm workers. In Karnataka, Gujarat, and Chhattisgarh, the agricultural workforce is smaller, with a shift towards non-agricultural employment, indicating economic diversification. At the other end of the spectrum, Jammu and Kashmir

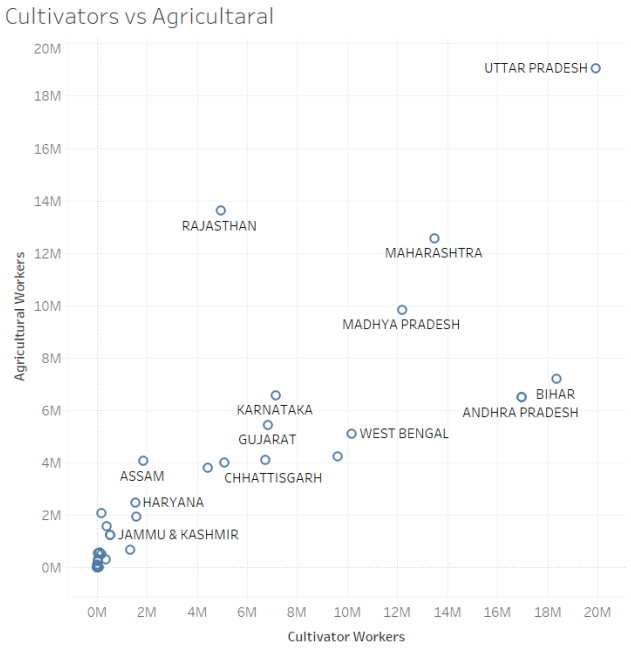

Fig. 35. Statewise Worker Category: Agricultaral vs Cultivator

and Haryana show the lowest numbers in both categories, highlighting their limited reliance on agriculture compared to other states.

Member-wise Contributions

The tasks were collaboratively discussed and agreed upon during an initial meeting. For the visualizations, dashboards, and narratives, the responsibilities were divided as follows:

- **Task 1:** Aditya Priyadarshi
- **Task 2:** Varnit Mittal
- **Task 3:** Ananthakrishna K

We combined our individual findings and analyses to finalize the project, including the report and video demonstration.

References

[1] https://www.kaggle.com/datasets/danofer/india-census
[2] National Urban Housing and Habitat Policy 2007
[3] Rent Control Act
[4] Wikipedia - India Census 2011
[5] Why does India still use and trade asbestos?
[6] Analysis of Female Workforce Participation Rate in the North-Eastern Region of India