India Demographics explorer
CS-661

Group-14

Aditya Anand
Abdul ahad
Tanish Bansal
Aurav pratap Singh

kartikey tomar
vibha narayan
ananya pandey
ruchika raj

Inspiration and creativity

"India's diversity is not just cultural—it's deeply demographic.
With over 1.2 billion people, understanding patterns in education, employment, and social structure is both a challenge and a necessity.
Through CensusScope, we reimagine static census data as a dynamic, interactive visual experience.
This project bridges data science and storytelling to make national insights accessible to all."

data cleaning and filter

01

..Merged Multiple Data Sources:
Compiled and standardized data from various Excel sheets into a single consolidated CSV file, ensuring uniform column headers, data types, and schema alignment across all sources.

02

Feature Selection & Pruning:
Removed irrelevant or redundant features/columns based on domain relevance, missing value thresholds, and correlation analysis to reduce noise and improve downstream processing..

03

.Value Normalization & Data Type Correction:
Transformed columns with inconsistent or raw numerical entries into meaningful formats—such as converting absolute numbers to percentages, scaling large values, and correcting data types (e.g., string to float/date).

04

Standardized Categorical Labels:
Unified inconsistent category entries (e.g., "M", "male", "MALE" → "Male"), corrected spelling variations, and applied consistent casing to ensure clean and reliable categorical data.

Overall Layout

State Analysis

Displays an interactive choropleth map of Indian states, where data metrics are visualized through a dynamic color gradient.
Users can click on any state to drill down into region-specific insights and initiate deeper analysis at the sub-state level.

District Analysis Section

Upon selecting a state, a district-wise choropleth map is generatedto provide granular insights.

The system automatically highlights the top 5 and bottom 5 districts based on the selected attribute, enabling quick identification of performance extremes.

Comparison Analysis

Enables side-by-side comparison of two states based on critical socio-economic indicators such as literacy rate, employment level, and age group distribution.
Helps in benchmarking and drawing meaningful inter-state comparisons.

State comparison

Exploring creativity

India chlorpeth map

shows distribution of that attribute across various states of india.

distribution summary

Visualizes the distribution of a selected attribute (e.g., literacy, employment) across Indian states using a box plot.
Highlights key statistics: mean, median, quartiles (Q1, Q3), and outliers for comparative insigh

key insights

tells summary about selected attribute like top performer, national average, performance gap etc.

state ranking

ranking different states according to different attributes

top-7 states

shows a pie-chart telling top-7 states of that attribute and average for that attribute.

correlation heatmap

shows a heatmap showing selected attributes relativity to other attributes.

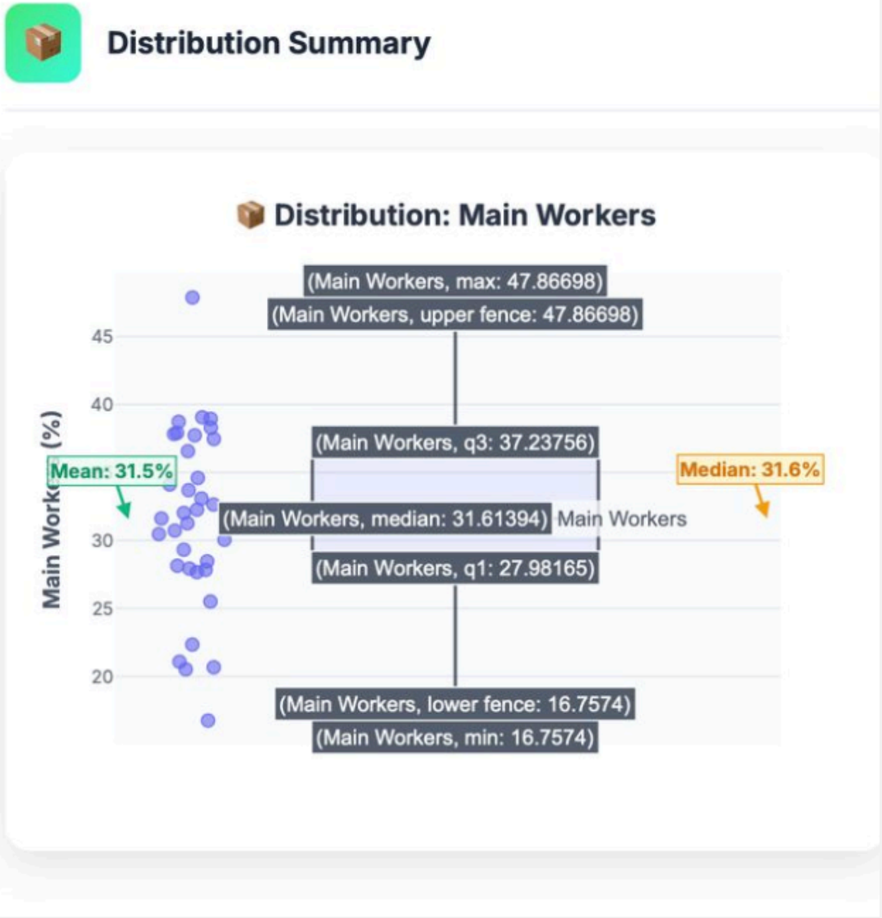
Box Plot Distribution

Visualizes the distribution of a selected attribute (e.g., literacy, employment) across Indian states using a box plot.

Highlights key statistics: mean, median, quartiles (Q1, Q3), and outliers for comparative insight.

Uses jittered data points and color-coded annotations to enhance interpretability.

Automatically updates based on attribute selection from the dropdown



Mathematical Formulas

Quartiles

$Q_1 = 25\text{th percentile}$

$Q_3 = 75\text{th percentile}$

Interquartile Range (IQR)

$\text{IQR} = Q_3 - Q_1$

Outlier Boundaries

Lower Bound: $Q_1 - 1.5 \times \text{IQR}$

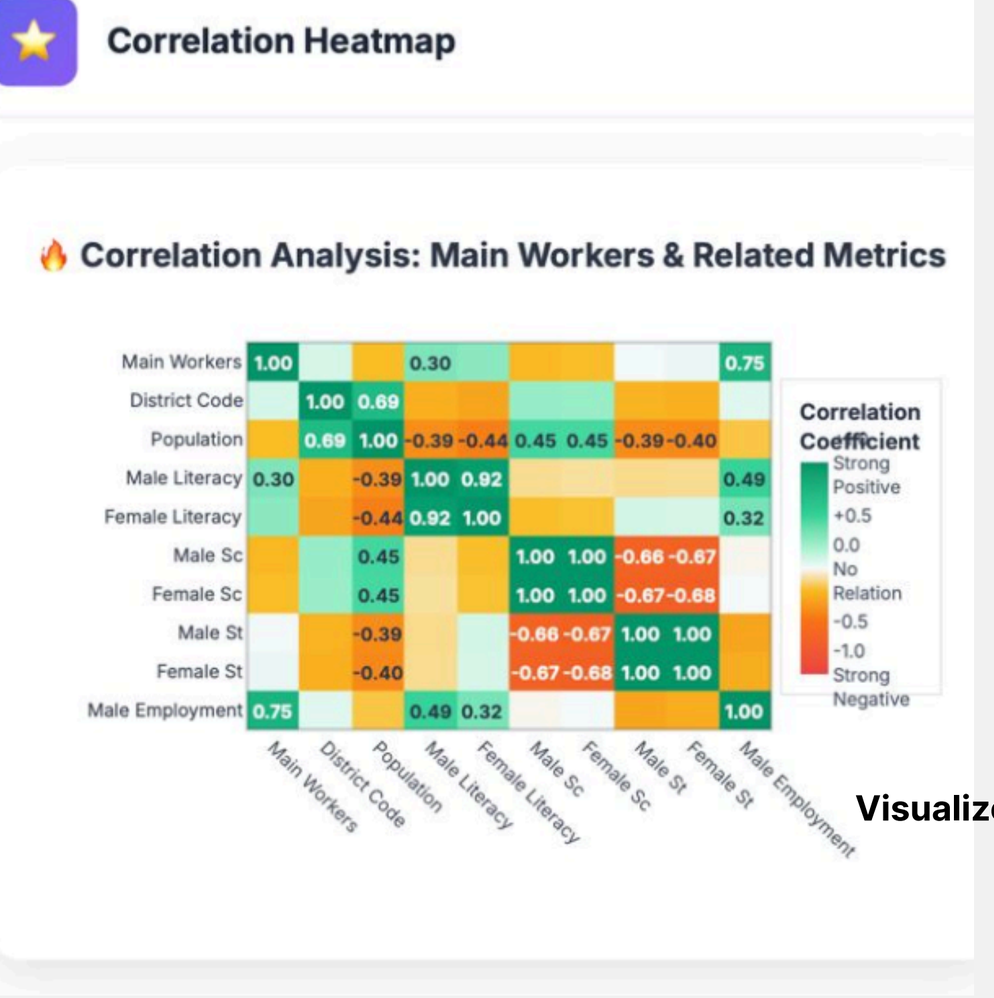
Upper Bound: $Q_3 + 1.5 \times \text{IQR}$

Mean

$$\mu = \frac{1}{n} \sum_{i=1}^{m} x_i$$

Standard Deviation

$$\sigma = \sqrt{\frac{1}{n}(x_i - \mu)}$$



Enhanced Correlation Heatmap

Visualizes pairwise correlations between selected demographic metrics using a color-coded matrix.

Numerical values are overlaid on each cell, emphasizing significant correlations (| r | >0.3 | r | >0.3) and diagonal values for clarity.

Custom color scale highlights the strength and direction of relationships (negative → red, positive → green).

Responsive and well-styled layout improves readability with rotated labels, consistent font, and interactive hover info.

district-analysis

district map

shows chloropeth map of state district-wise according to selected attribute.

district rankings

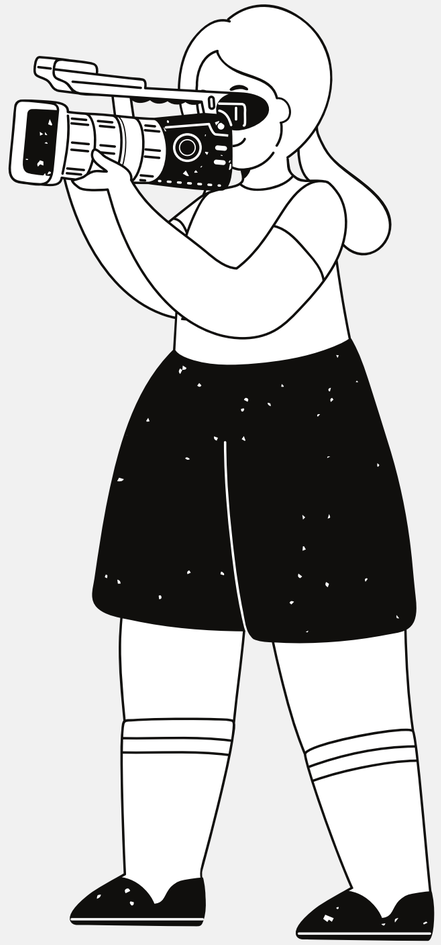
ranks different districts of state according to selected attribute in chloropeth colors.

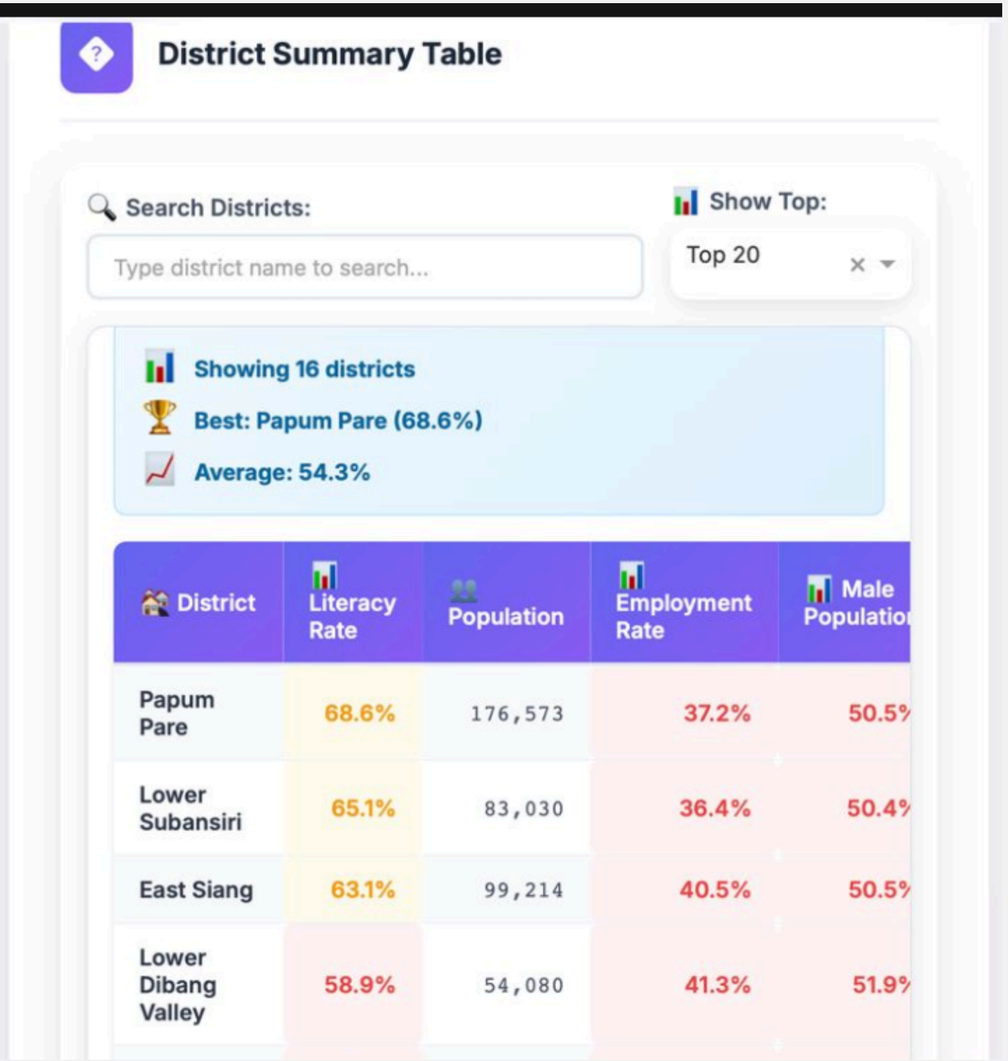
performance matrix

makes a scatter-plot of selected attribute vs population.

district summary table

shows whole summary of state according to selected attribute and its related attribute,average,best district according to that attribute.





Sorting by Metric (Ranking Logic)

1. Ranks districts from highest to lowest performers

Summary Statistics Calculated

2. avg_value = table_data[selected_attribute].mean,

$$\mu = \frac{1}{n}\sum_{i=1} x_i$$

best_district = table_data.iloc[0][District name]. best_value = table_data iloc[0][selected_attribute]

Performance Categorization with Thresholds

3. 🔴 Low (<60%) 🟡 Moderate (60–80%)
🟢 High (>80%)

This dynamic table summarizes district-level performance across key metrics such as literacy, employment, and gender distribution.

Users can search, filter, and sort districts based on selected attributes with intuitive color-coded indicators.

Performance levels are visually enhanced using icons and colored backgrounds for quick interpretation.

Additional statistics like average value and best-performing district provide instant context for data-driven decisions.

comparison window

Exploring creativity

multi-dimensional radar chart

radar chart enables multi-dimensional comparison of selected attribute enabling comparing different states at once.

bar chart comparison

compare different selected states for a selected attribute.

development pathway analysis

This visualization benchmarks states by comparing their current performance with their potential

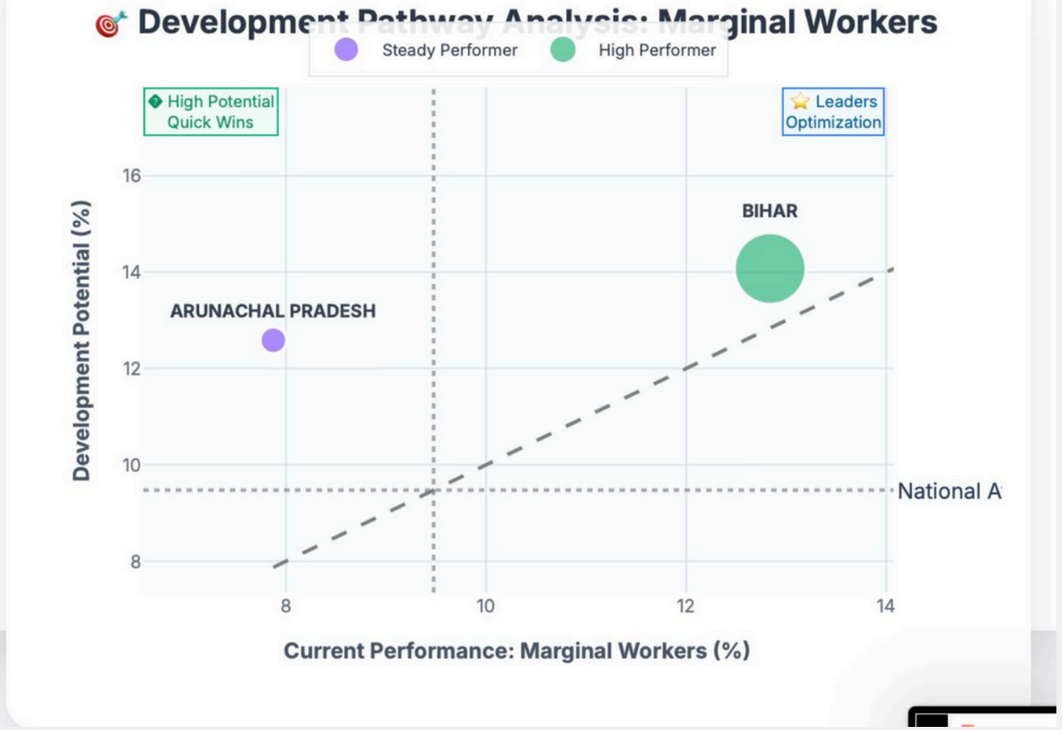
comparison data table

compares and ranks different selected states for different dimensions of that selected attribute.

Key insights

tells summary about selected attribute like top performer, performance leader etc.

Development Pathway Analysis



Mathematical Formulations Behind Pathway Analysis or Tables

Development Pathway Analysis

1. National Average $\mu = \frac{1}{n}\sum_{i=1} x_i$
2. Development Potential (Custom Rule)

$$\text{Potential}_i = \begin{cases} x_i + 0.7\,(\text{Top}10_i - x_i), & \text{if } x_i < \text{Top}_{10_i} \\ 1.05 \cdot x_i & \text{otherwise} \end{cases}$$

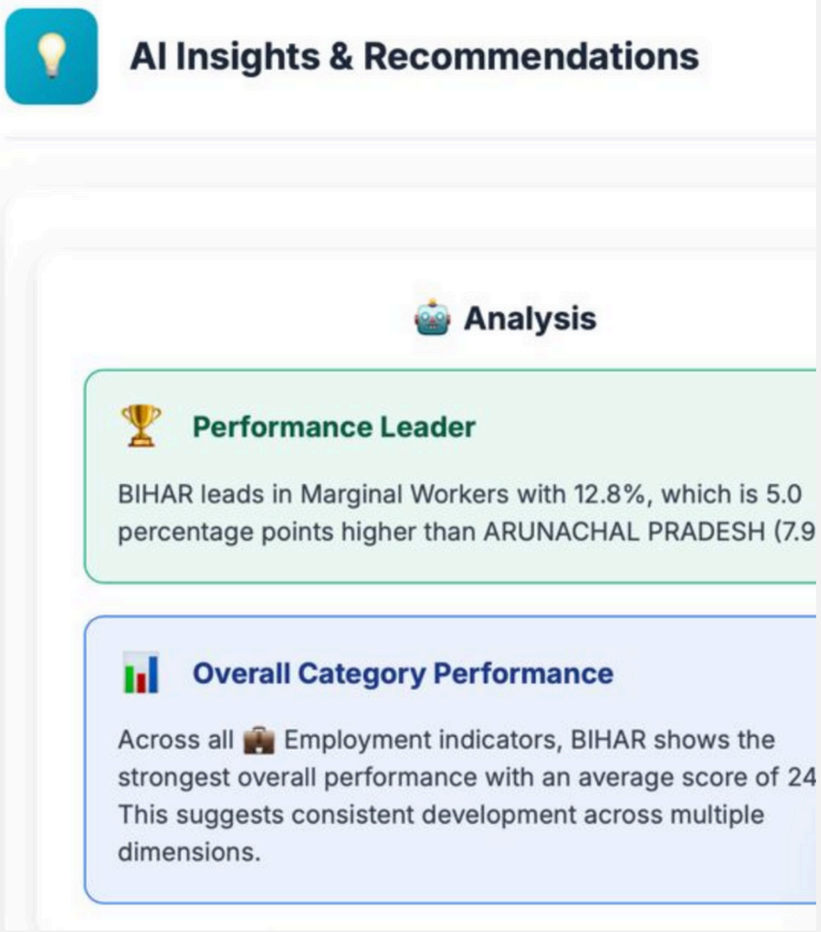
3. Improvement Gap $\Delta_i = \text{Potent}_{il} - x_i$
4. Categorization Heuristics $\text{if } x_i > 1.2\,\mu \quad x_i < 0.8\,\mu$

Comparison Data Table

1. Ranking (min method) $\text{Ran}k_i = \text{of } x_i \text{ in sorted list}$
2. Styling Logic Based on Rank
   - # 1
   - # 2
   - # 3
   - Others
   - # 1
   - Others

This visualization benchmarks states by comparing their current performance with their potential.

- **Bubble size represents population, while position highlights gaps and opportunities for growth.**
- **States are categorized into stages like High Performer, Growth Leader, or High Potential based on performance vs. national average.**
- **Diagonal and reference lines help visualize how close a state is to its optimized potential.**



Logic & Methodologies Behind AI Insight Generation

1. Performance Leader Identification

$\text{Gap} = \max(x_i) - \min(x_i)$

Highlights

Overall Category Score

$Avg_{category} = \frac{1}{k} S_{ij}$ (for each)

X (for each state)

Development Consistency

$\text{Range}_i = \max(x_{ij}) - \min(x_{ij})$

(within a state)

Strategic Recommendation Logic

Uses keyword-based heuristics on attribute name

Best Practices Identification

Triggers if ≥ 3 states selected

Extracts top performer

insights highlight top-performing states, average category scores, and development consistency.

- **Comparative gap analysis reveals disparities between highest and lowest states for any selected attribute.**
- **The system provides strategic suggestions (e.g., focus areas, best practices) based on category and performance.**
- **Visual cards enhance interpretability using emojis, color-coded alerts, and contextual explanations.**

Challenges

1.Handling High-Dimensional Data

- Challenge: The census dataset includes dozens of categories and hundreds of attributes, many with inconsistent formats or naming conventions in geojson.
- data was in string and float,int etc. changing numerical data percentage.

2. Dynamic Interactivity Across Components

- Coordinating multiple callbacks (state dropdowns, attribute selectors, visualizations, insights) without cyclic dependencies or triggering issues.
- Avoided our first pop-up window model and changed our layout to select.
- fitting map in that card and choosing chloropeth.

3.Evolution and change of map through year

like formation of telangana and change of laddakh (previously part of JK) now UT.

.

4.Visualization Clarity & Responsiveness

- Making sure charts (e.g., radar, box, correlation heatmaps) are readable, well-aligned, and informative across screen sizes.

5.Generating Textual Insights

- Crafting logic to compare states and extract textual recommendations using only numerical data.

Thank you very much!