Anaylsis and Market Segmentation

On

India Primary healthcare based on Geographic and workforce data

by

Prajakta Khairnar

Sumit kushwaha

Github links for code

Click Here For Geographic data

Click Here for workforce data

Problem Statement

India's rural healthcare system, comprising Sub Centres (SCs), Primary Health Centres (PHCs), and Community Health Centres (CHCs), faces significant challenges in workforce availability and infrastructure distribution. Using 2017 state-level data on healthcare facilities, personnel, and population coverage, this project aims to analyze disparities in service delivery across states. Key indicators such as shortfall, vacancy rate, and population served per facility help identify underserved regions and overburdened health units. Geographic metrics like villages and rural population per center reveal structural imbalances. The analysis further segments states based on demographic, economic, and geographic characteristics to uncover systemic gaps. High average travel distances to CHCs serve as a proxy for tech adoption readiness, highlighting regions where digital health solutions could be impactful. This integrated approach supports data-driven policymaking, efficient resource allocation, and targeted interventions to strengthen rural healthcare delivery in India.

Data Collection

Links for data extraction:

for geography

https://www.kaggle.com/datasets/webaccess/india-primary-health-care-data?select=rural-area-covered-centre_2017.csv

https://www.kaggle.com/datasets/webaccess/india-primary-health-care-data?select=rural-population-centre_2017.csv

https://www.kaggle.com/datasets/webaccess/india-primary-health-care-data?select=villg-coveredby-centre_2017.csv

for healthcare workforce

https://www.kaggle.com/datasets/webaccess/india-primary-health-care-data?select=allo-doc-PHCS_2017.csv

https://www.kaggle.com/datasets/webaccess/india-primary-health-care-data?select=nursing-staff-PHCS-CHCS_2017.csv

https://www.kaggle.com/datasets/webaccess/india-primary-health-care-data?select=pharmacists-PHCS-CHCS_2017.csv

https://www.kaggle.com/datasets/webaccess/india-primary-health-care-data?select=physician-CHCS_2017.csv

https://www.kaggle.com/datasets/webaccess/india-primary-health-care-data?select=radiographers-CHCS_2017.csv

https://www.kaggle.com/datasets/webaccess/india-primary-health-care-data?select=rural-area-covered-centre_2017.csv

https://www.kaggle.com/datasets/webaccess/india-primary-health-care-data?select=assistant-female-PHCS_2017.csv

https://www.kaggle.com/datasets/webaccess/india-primary-health-care-data?select=assistant-male-PHCS_2017.csv

https://www.kaggle.com/datasets/webaccess/india-primary-health-care-data?select=worker-female-subcen_2017.csv

https://www.kaggle.com/datasets/webaccess/india-primary-health-care-data?select=worker-male-subcen_2017.csv

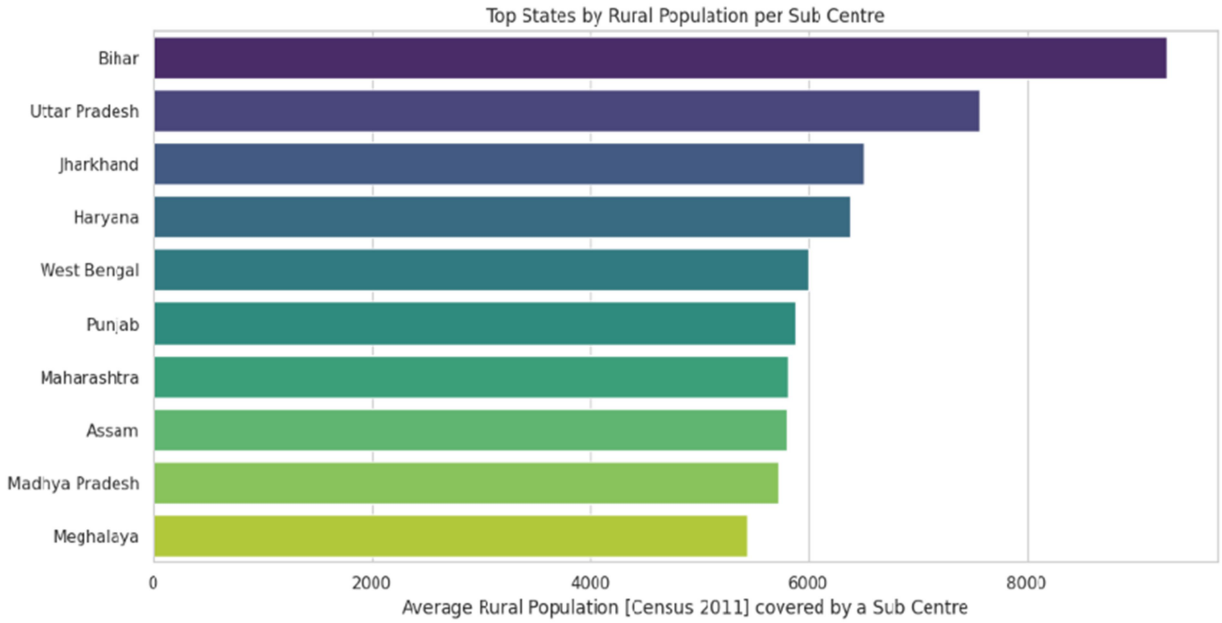
1.Geographic Based Healthcare Market Segmentation Report

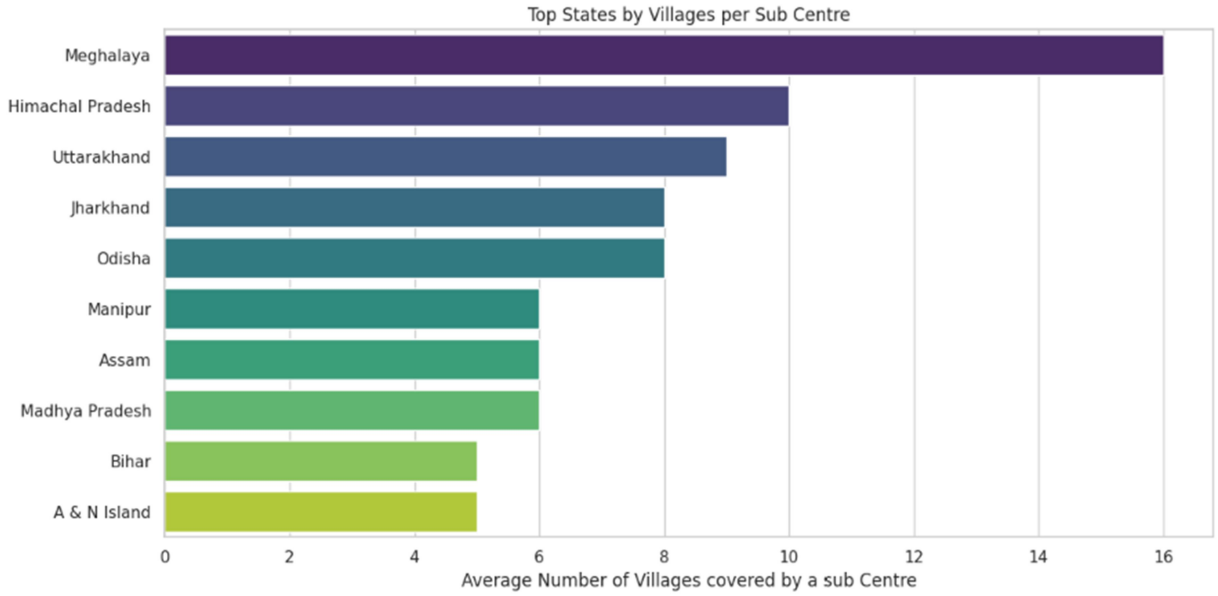
Segment 1: Demographic & Firmographic

To understand how rural populations and villages are distributed across healthcare facilities, indicating service reach and potential bottlenecks.

Key Analyses:

- **Average Rural Population per Sub Centre**
- **Average Number of Villages per Sub Centre**





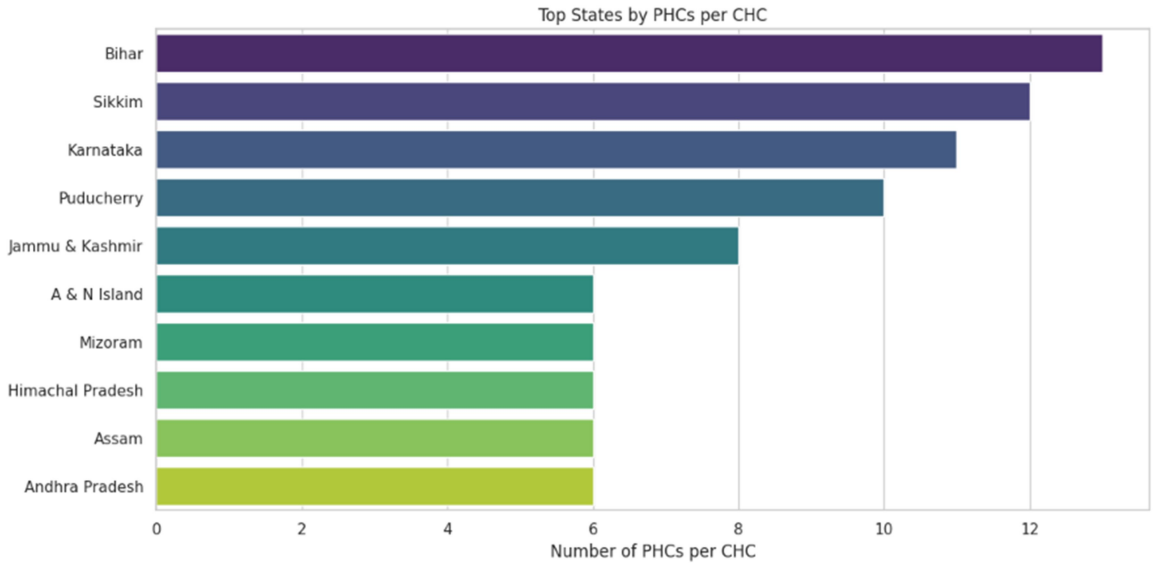
This segmentation highlights states where Sub Centres cover disproportionately large rural populations and village counts, indicating strain on basic healthcare access. High values suggest overburdened infrastructure and a need for more localized service points. These insights support planning for equitable distribution of Sub Centres.

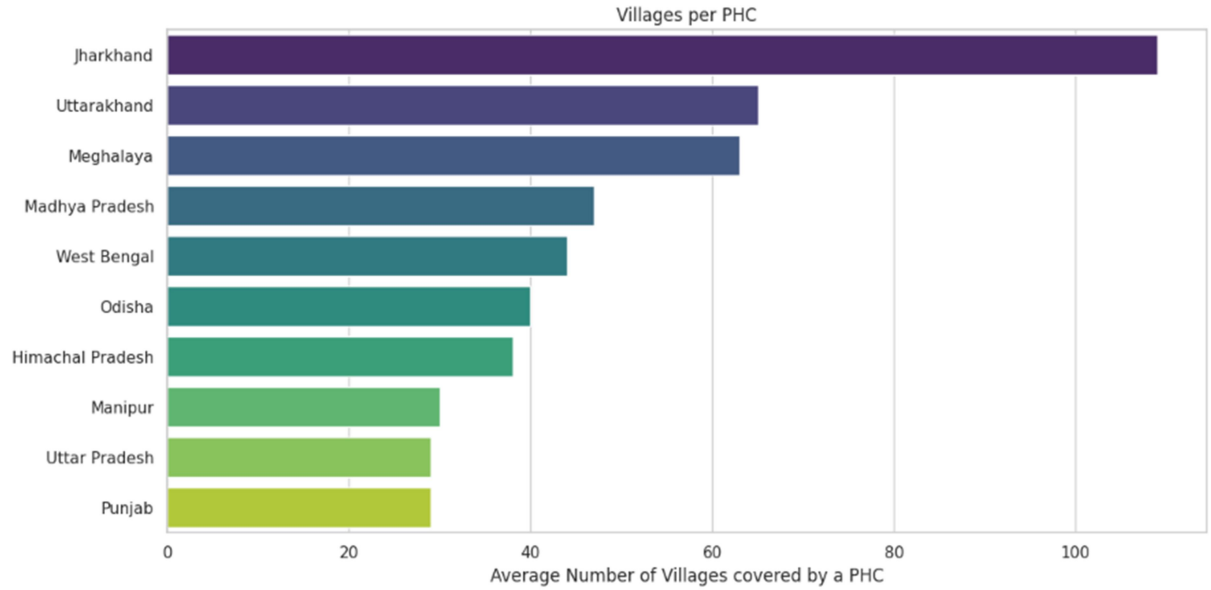
Segment 2: Service & Role-Based Distribution

To examine how healthcare services are layered structurally from Sub Centres to PHCs to CHCs and how they interconnect in coverage.

Key Analyses:

- **Number of PHCs per CHC**
- **Villages covered per PHC**





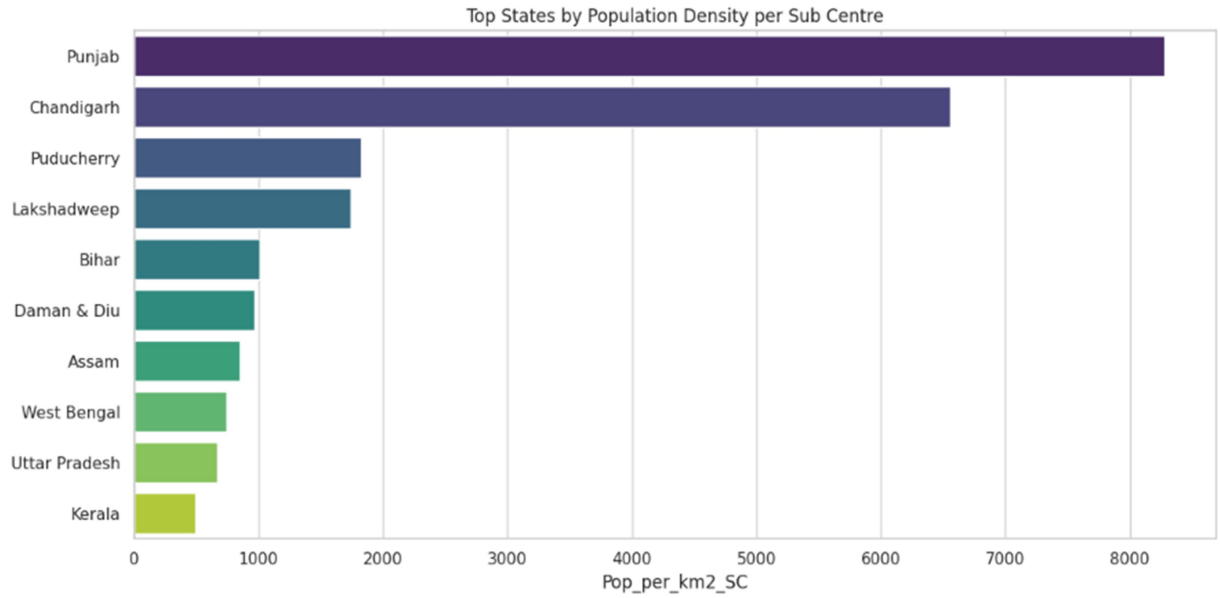
This segmentation reveals the structural distribution of Primary Health Centres (PHCs) and Community Health Centres (CHCs) across states. A high PHC-to-CHC ratio or large village coverage per PHC indicates potential service overload at the primary care level. Such imbalances highlight areas needing better infrastructure or role redistribution.

Segment 3: Resource & Economic Segmentation

To evaluate the **efficiency of population coverage** in terms of geography population served per square kilometer of Sub Centre area.

Key Analysis:

- **Population Density per Sq. Km per Sub Centre** (Pop_per_km2_SC)



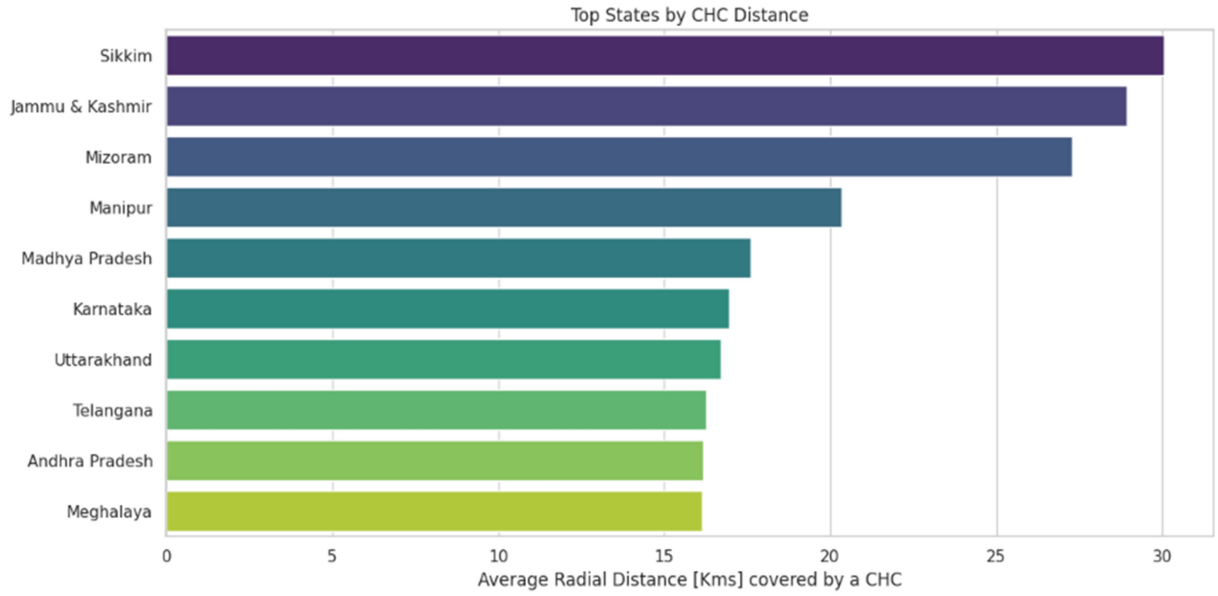
This segmentation measures the population density served by each Sub Centre, highlighting states where healthcare resources are stretched over densely populated rural areas. High density values suggest increased demand and pressure on limited infrastructure. These insights help target investments in high-need, high-density regions.

Segment 4: Behavioral & Technology Adoption Segmentation

To use **radial distance coverage** as a proxy for both travel burden and potential areas for **technology-driven interventions** (e.g., telemedicine, eHealth).

Key Analysis:

- **Average Radial Distance Covered by CHCs**



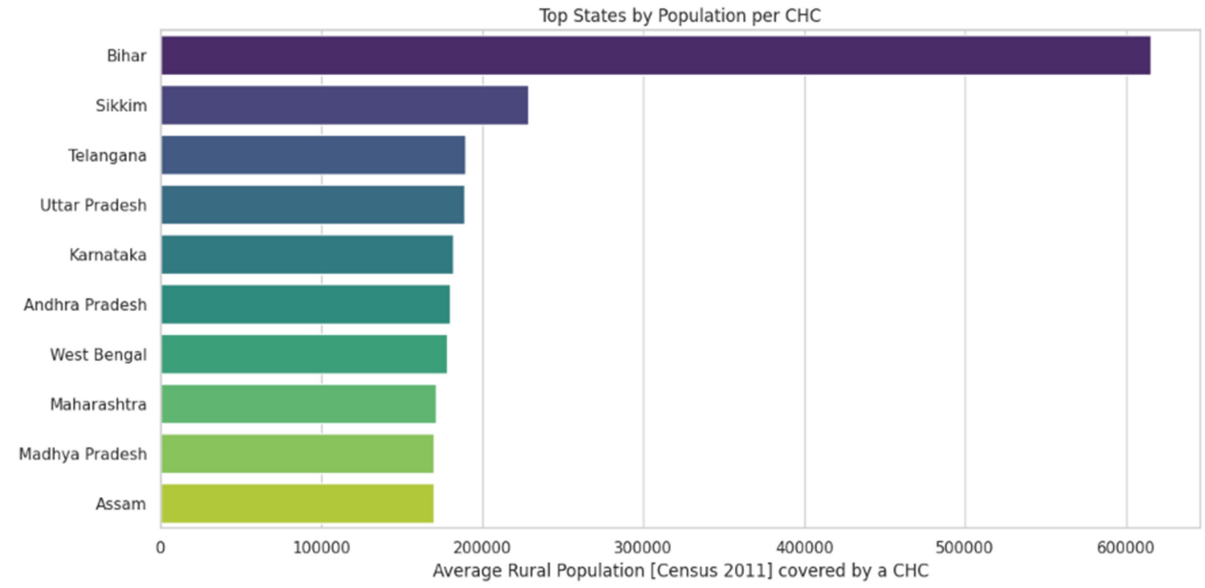
This segmentation uses radial distance covered by CHCs as a proxy for accessibility and potential technology adoption. States with larger average distances may face challenges in timely service delivery, suggesting a need for mobile health solutions or digital interventions. It reflects infrastructural dispersion and access barriers.

Segment 5: Service Delivery & Public Health Impact

To estimate **potential health outcomes and infrastructure strain** by analyzing the rural population load per CHC.

Key Analysis:

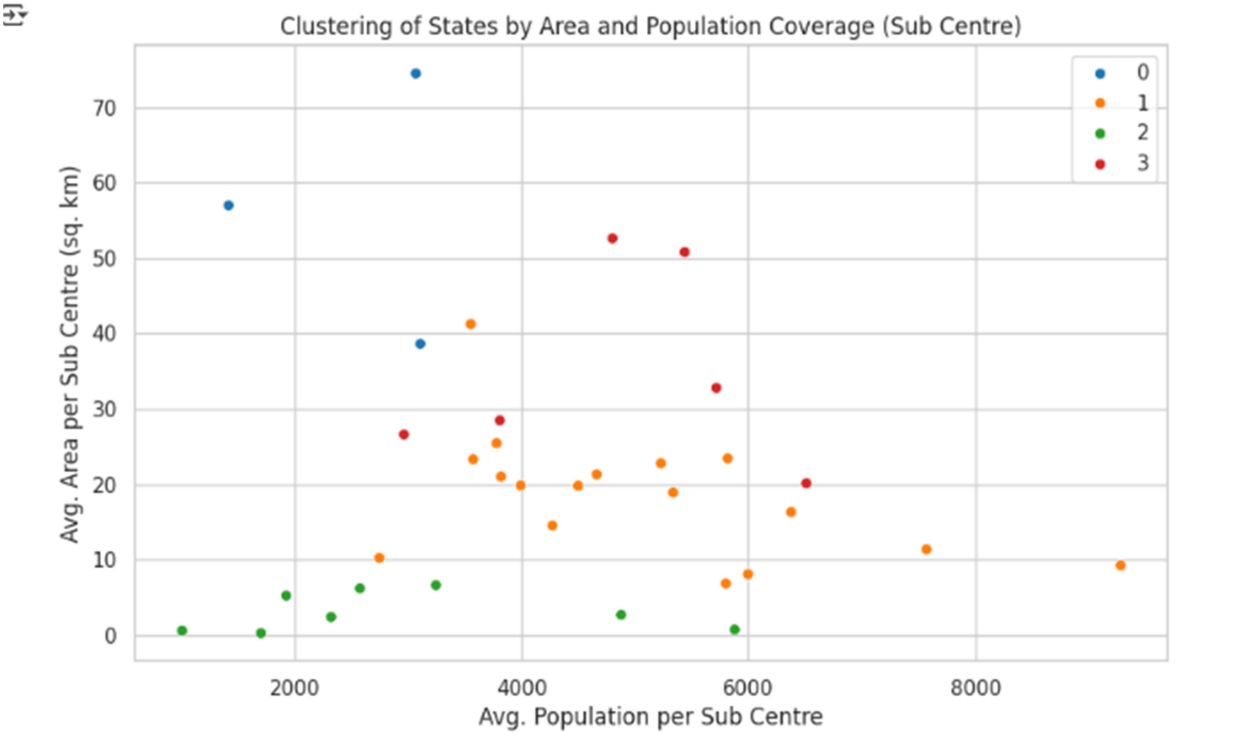
- **Average Rural Population Covered by a CHC**



This segmentation highlights the population served by each Community Health Centre (CHC), indicating states with high pressure on CHC resources. States with higher population coverage per CHC may experience delays in service delivery and compromised healthcare quality. This underscores the need for expanding CHC capacity or redistributing resources for better service coverage.

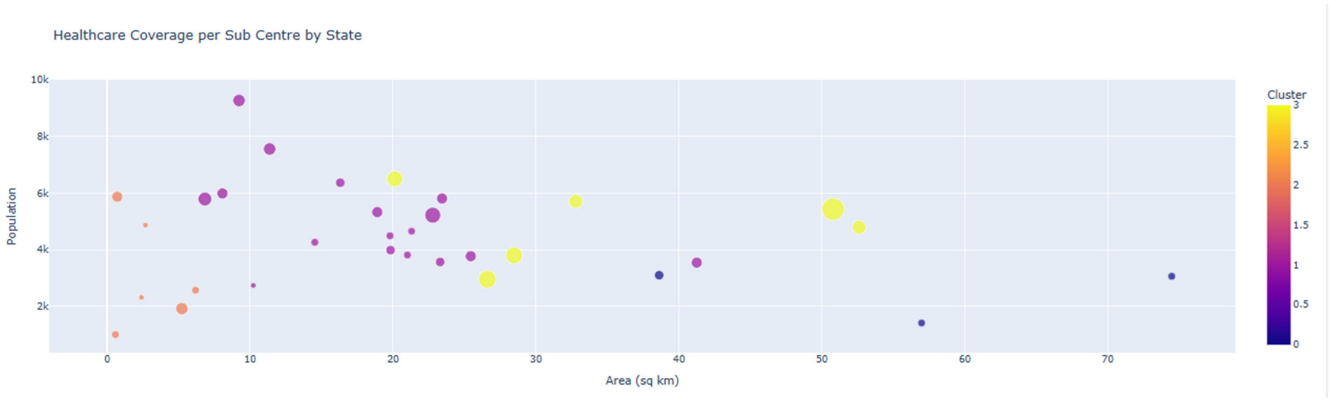
Segment Extraction by using k-means clustering

The scatter plot visualizes how Indian states are clustered based on the average rural population and area covered by Sub Centres. Each point represents a state, and colors indicate different clusters formed using unsupervised learning. States grouped together share similar geographic healthcare coverage patterns. High values on both axes suggest overburdened Sub Centres serving large populations across wide areas. This visualization helps identify regions that may require targeted infrastructure expansion or resource reallocation.



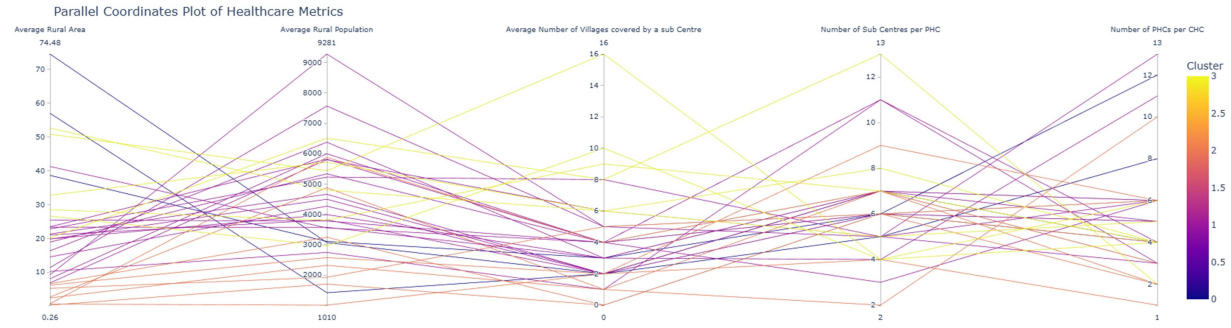
Bubble chart (Area vs Population)

The bubble chart illustrates the geographic and demographic burden on Sub Centres across Indian states. Each bubble represents a state, with its position showing average area (x-axis) and population (y-axis) covered per Sub Centre. The bubble size reflects the average number of villages served, and colors denote cluster groupings. Larger and higher-positioned bubbles indicate states where Sub Centres are overextended in both reach and responsibility. This visualization highlights disparities in rural health infrastructure and aids in identifying states needing urgent capacity upgrades.



Parallel Coordinates Plot

The parallel coordinates plot compares multiple healthcare infrastructure metrics across Indian states, with lines colored by cluster. Each line represents a state, and its position across axes reflects values for rural area, population, villages per Sub Centre, and service ratios (SC/PHC, PHC/CHC). This visualization highlights patterns and outliers across multiple dimensions simultaneously. States with similar infrastructure profiles cluster together, making it easier to spot systemic gaps or overburdened service structures. It's a powerful tool for multivariate comparison in rural healthcare planning.



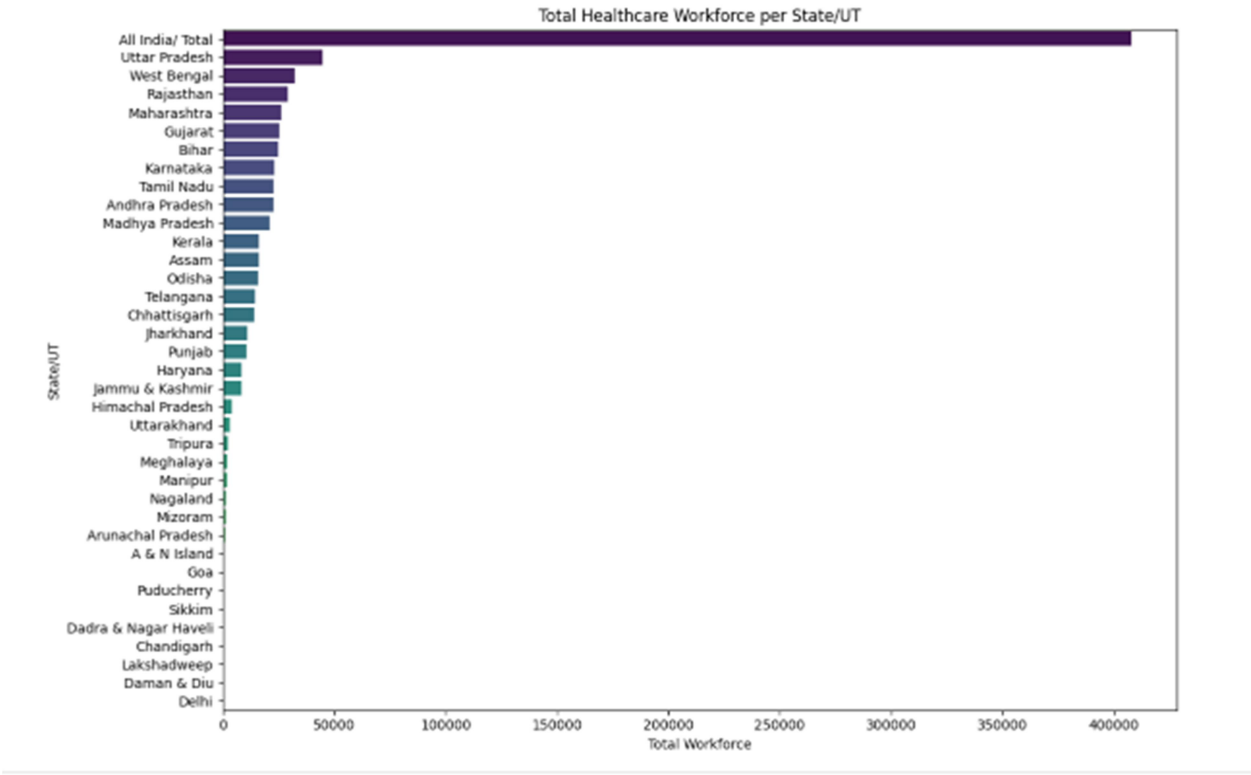
2. Healthcare Workforce Market Segmentation Report

Segment 1: Demographic & Firmographic

To understand workforce distribution across Indian states/UTs at a macro (firmographic) level using headcount data.

Key Analysis:

- Computed total workforce per state by summing all roles from the **In_Position** column.
- Generated a **barplot** of **Total_Workforce** per state to compare healthcare staffing levels.

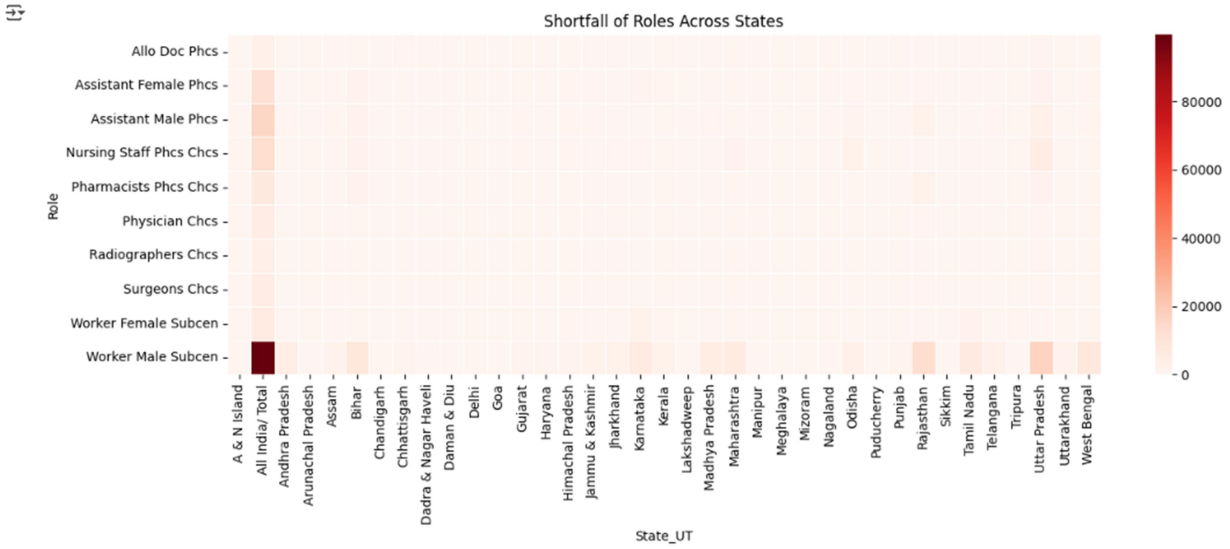


The bar plot shows the total healthcare workforce across Indian states, sorted from highest to lowest. It highlights which states have the largest staffing levels, offering insight into resource concentration. This aids in identifying workforce distribution gaps across regions.

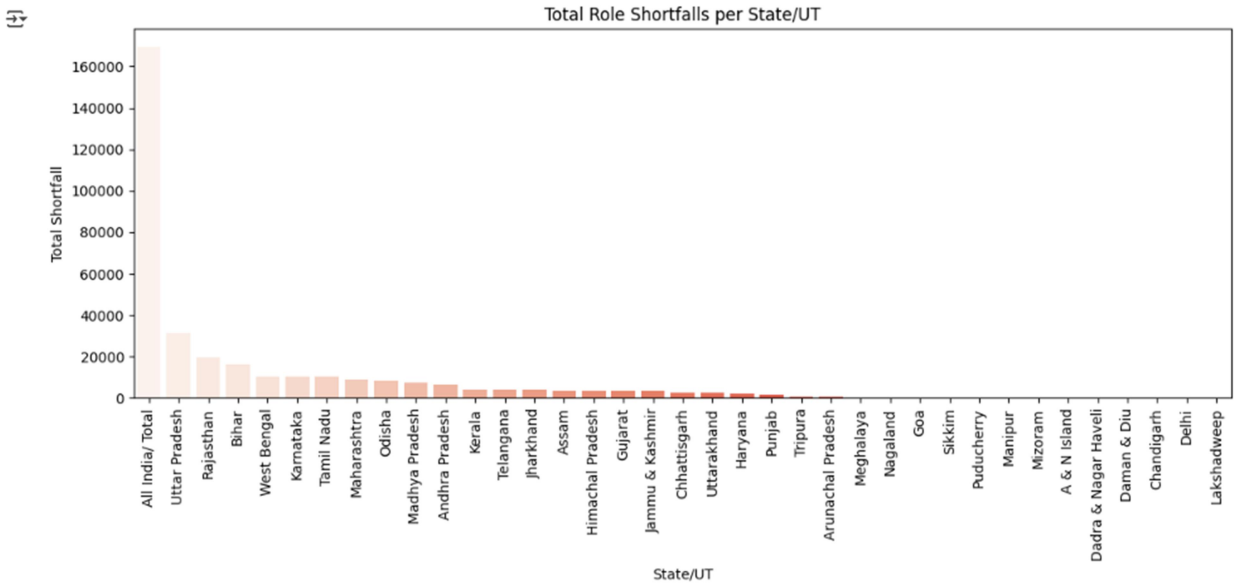
Segment 2: Service/Role Distribution.

Key Analysis:

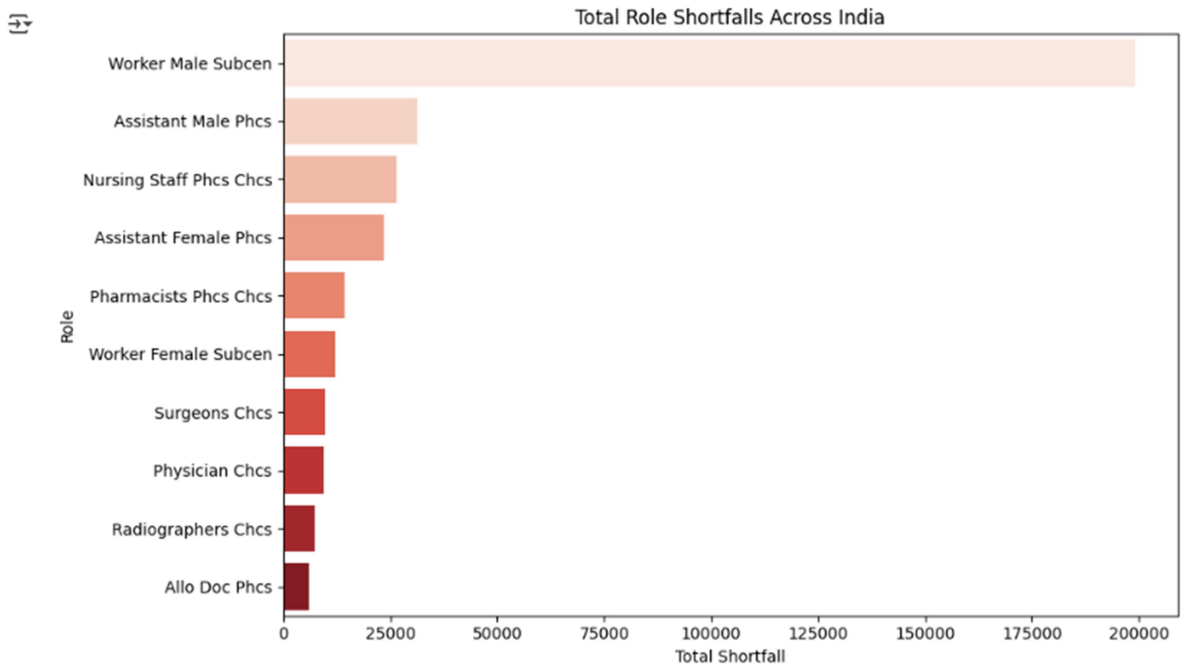
- Created a pivot table of Shortfall by Role and State_UT.
- Generated both a **heatmap** and **barplots** (aggregated by role and state) to visualize shortfall patterns.



The heatmap visualizes the shortfall of healthcare roles across Indian states, with darker shades indicating higher shortages. It helps identify which roles are most understaffed in each state. This supports targeted workforce planning and role-specific interventions.



The bar plot shows the total shortfall of healthcare roles in each Indian state, sorted from highest to lowest. States with the tallest bars face the most significant staffing gaps. This visualization highlights where urgent workforce reinforcement is needed.



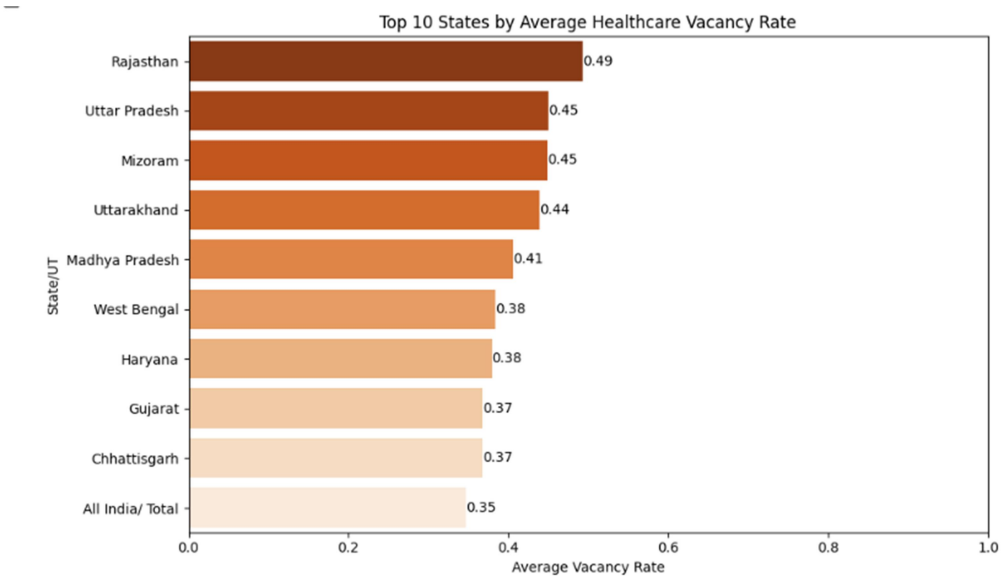
The bar plot illustrates the total shortfall for each healthcare role across India, ranked from highest to lowest. Roles with the longest bars are the most understaffed nationwide. This helps prioritize recruitment and training efforts for critical roles.

Segment 3: Financial & Economic (Vacancy Rates)

Objective:
To assess staffing efficiency using **vacancy rate** as a proxy for fiscal allocation vs actual hiring.

Key Analysis:

- Defined Vacancy_Rate as (Sanctioned - In_Position) / Sanctioned.
- Averaged vacancy rates per state and visualized using a **horizontal barplot**.



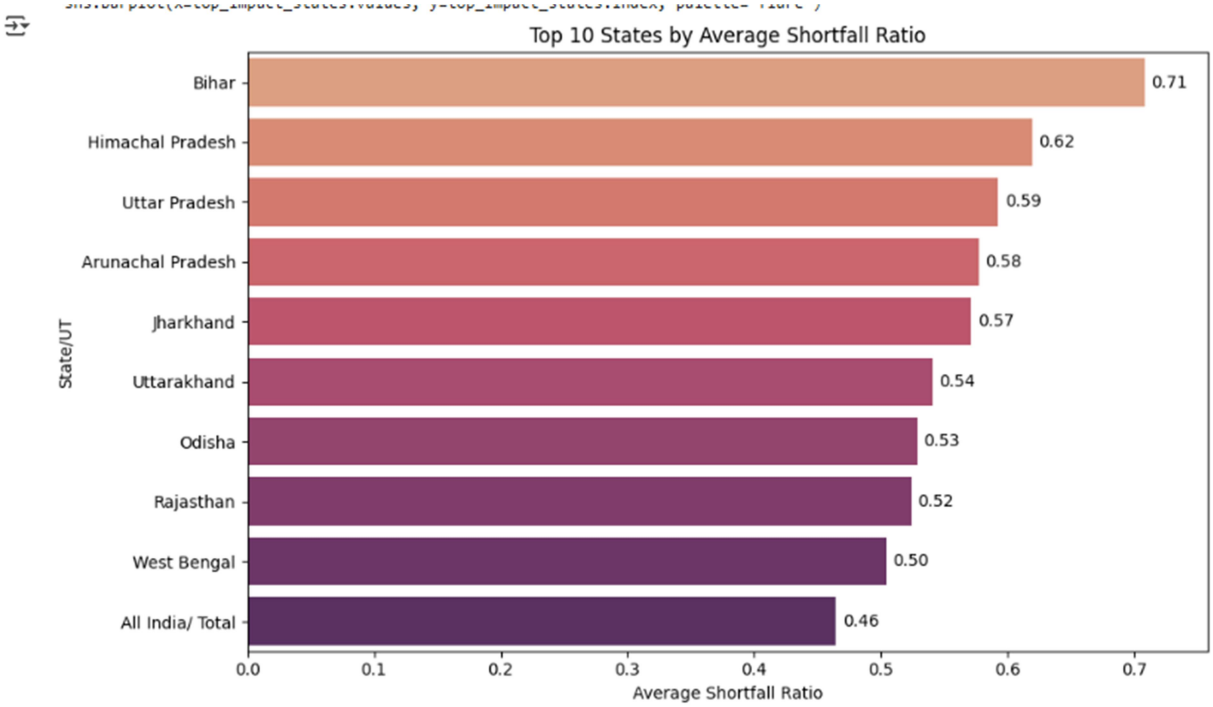
The bar plot displays the top 10 Indian states with the highest average healthcare vacancy rates. It highlights regions where a large proportion of sanctioned posts remain unfilled. This helps identify states facing critical staffing shortages and requiring urgent workforce interventions.

Segment 4: Tech Adoption (Proxy via Low Staffing)

Objective:
To indirectly infer tech dependence or potential for tech adoption by examining **Shortfall Ratios**, especially in roles that could be digitized or automated.

Key Analysis:

- Computed Shortfall_Ratio = Shortfall / (Required + 1) to normalize shortfalls across states.
- Ranked states based on average shortfall ratios.



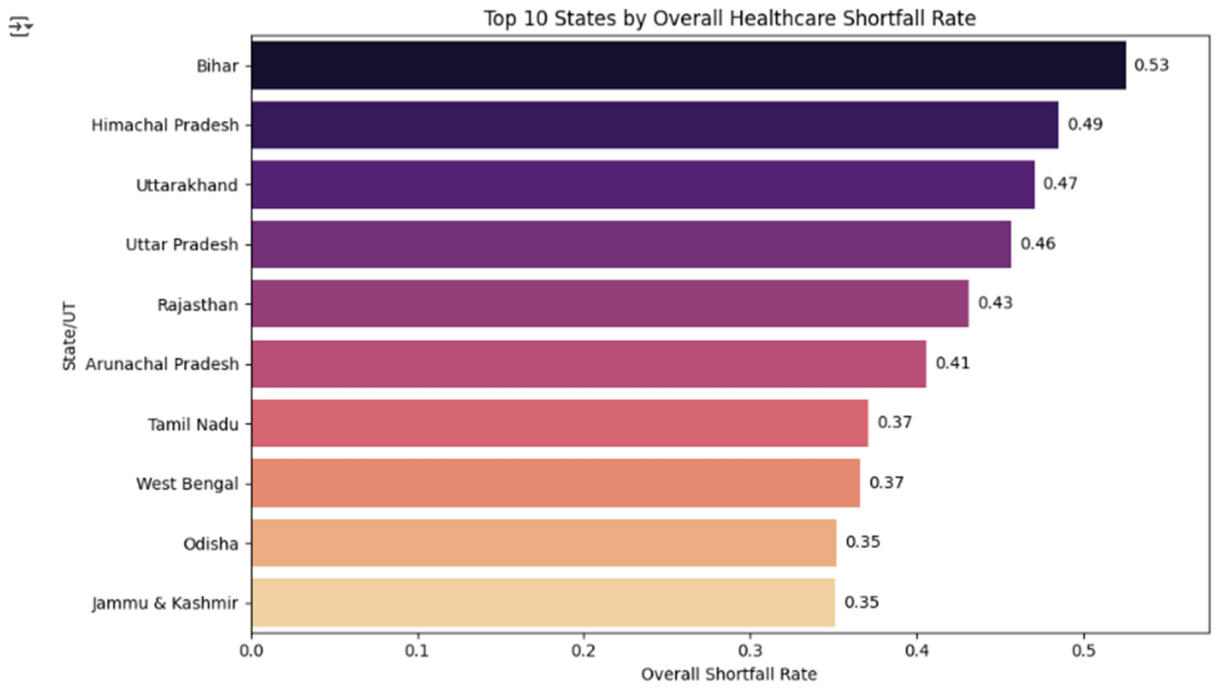
The bar plot highlights the top 10 Indian states with the highest average shortfall ratios, indicating the proportion of required positions that are unfilled. States with higher ratios are experiencing the greatest service delivery gaps. This visualization supports prioritizing healthcare resource allocation to the most underserved regions.

Segment 5: Market Need (Shortfall vs Required)

To quantify and rank **market demand** for workforce deployment and investments using total shortfall and required values.

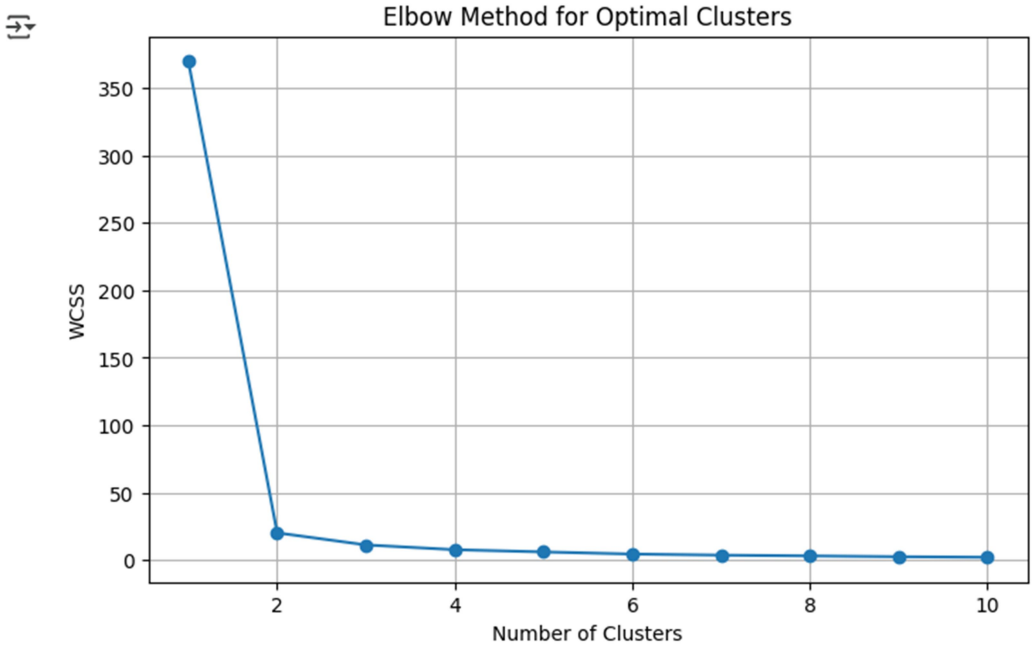
Key Analysis:

- Grouped by state and summed Required, Sanctioned, In_Position, and Shortfall.
- Computed an **overall Shortfall Rate** = Shortfall / (Required + 1) and ranked states.
- Visualized top 10 states via barplot.



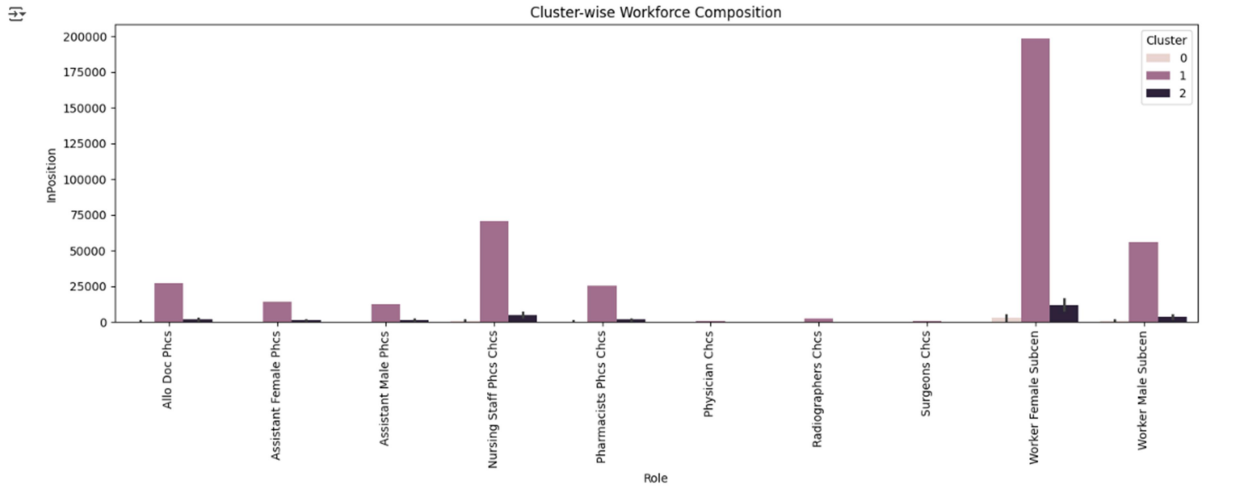
The bar plot shows the top 10 states with the highest overall healthcare shortfall rates, based on total required versus actual staffing. Higher bars indicate states with more severe gaps in fulfilling healthcare staffing needs. This visualization aids in identifying critical regions for policy and resource intervention.

Clustering (K-Means for Segmentation)



The Elbow Method plot helps determine the optimal number of clusters for segmenting states based on role-wise in-position workforce data. The "elbow point" in the curve indicates where adding more clusters yields diminishing returns in reducing within-cluster variance (WCSS). This guides effective clustering for workforce distribution analysis.

Cluster Profile Analysis



The bar plot shows the distribution of in-position healthcare roles across different clusters of states. Each cluster represents a group of states with similar workforce compositions, and the plot compares role-wise staffing levels within them. This helps understand how role allocation varies across state clusters.