**Healthcare Staffing Segmentation Analysis Report**

This report provides a comprehensive, step-by-step summary of the market segmentation analysis performed on healthcare staffing data from the allo-doc-PHCS\_2017.csv dataset. The analysis focuses on staffing levels for doctors, nurses, and surgeons across Primary Health Centers (PHCs) in Indian states and Union Territories, aiming to identify distinct staffing profiles and geographic disparities to inform targeted workforce strategies.

**1. Objective**

The goal was to segment Indian states and Union Territories based on healthcare staffing metrics and PHC area coverage, identifying regions with staffing shortages, imbalances, or geographic challenges. This segmentation helps prioritize interventions to improve healthcare access and workforce distribution.

**2. Data Overview**

**Dataset Description**

* **Source**: allo-doc-PHCS\_2017.csv
* **Rows**: 36 (after excluding 'All India/ Total')
* **Columns**: 10, including:
  + State/ UT: Name of the state or Union Territory.
  + Required - [R], Sanctioned - [S], In Position - [P], Vacant - [S-P], Shortfall - [R-P]: Staffing metrics for doctors.
  + Notes and additional columns for nurses and surgeons (assumed to be similarly structured in the full dataset).
* **Additional Data**: PHC area coverage (derived or provided) to assess geographic workload.

**Initial Exploration**

* The dataset was loaded using pandas, and its structure was verified:
* df1 = pd.read\_csv('allo-doc-PHCS\_2017.csv', delimiter=',', na\_values='NA', nrows=1000)
* df1 = df1[df1['State/ UT'] != 'All India/ Total']
* print(f'There are {df1.shape[0]} rows and {df1.shape[1]} columns')

**Output**: 36 rows, 10 columns.

* A sample of the data was inspected to understand its structure:
* print(df1.head(5))

**Output**: Showed columns like State/ UT, Required - [R], Sanctioned - [S], In Position - [P], etc., with some missing values (e.g., NaN for sanctioned positions in Arunachal Pradesh).

**3. Methodology**

The analysis followed a structured approach, involving data preprocessing, feature engineering, clustering, visualization, and interpretation.

**Step 1: Data Preprocessing**

* **Handling Missing Values**:
  + Missing values (NaN) in columns like Sanctioned - [S], Vacant - [S-P], and Shortfall - [R-P] were addressed. Common strategies (inferred from the code) include:
    - Imputing with zeros for vacancy or shortfall where appropriate (e.g., if no sanctioned positions exist).
    - Excluding or flagging states with excessive missing data for certain analyses.
* **Data Cleaning**:
  + Removed the 'All India/ Total' row to focus on individual states/UTs.
  + Standardized column names (e.g., renamed State/ UT to State for consistency).

**Step 2: Feature Engineering**

To enable clustering, relevant features were derived or computed:

* **Staffing Ratios**:
  + Doc\_Staffing\_Ratio: Percentage of doctors in position relative to required ([P]/[R] \* 100).
  + Nurse\_Staffing\_Ratio: Similarly computed for nurses.
  + Surgeon\_Staffing\_Ratio: Similarly computed for surgeons.
* **Shortfall Percentages**:
  + Doc\_Shortfall\_Pct: Shortfall as a percentage of required doctors ([R-P]/[R] \* 100).
  + Nurse\_Shortfall\_Pct: Similarly computed for nurses.
  + Surgeon\_Shortfall\_Pct: Similarly computed for surgeons.
* **Vacancy Percentages**:
  + Doc\_Vacancy\_Pct: Vacancy as a percentage of sanctioned positions ([S-P]/[S] \* 100).
  + Nurse\_Vacancy\_Pct: Similarly computed for nurses.
  + Surgeon\_Vacancy\_Pct: Similarly computed for surgeons.
* **Geographic Feature**:
  + PHC\_Area\_Covered: Area (in square kilometers) covered by each PHC, reflecting geographic workload.

**Step 3: Data Standardization**

* Features were standardized to ensure equal weighting during clustering:
* from sklearn.preprocessing import StandardScaler
* scaler = StandardScaler()
* scaled\_data = scaler.fit\_transform(data[features])
* **Features Used**: Doc\_Shortfall\_Pct, Doc\_Vacancy\_Pct, Doc\_Staffing\_Ratio, Nurse\_Shortfall\_Pct, Nurse\_Vacancy\_Pct, Nurse\_Staffing\_Ratio, Surgeon\_Shortfall\_Pct, Surgeon\_Vacancy\_Pct, Surgeon\_Staffing\_Ratio.
* Standardization ensured that features with different scales (e.g., percentages vs. ratios) did not disproportionately influence the clustering.

**Step 4: Clustering**

* **Algorithm**: KMeans clustering was applied to segment states into distinct groups.
* from sklearn.cluster import KMeans
* kmeans = KMeans(n\_clusters=4, random\_state=42)
* data['Cluster'] = kmeans.fit\_predict(scaled\_data)
* **Number of Clusters**: Four clusters were chosen (likely based on elbow method or domain knowledge, though not explicitly shown in the code).
* **Cluster Naming**: Clusters were labeled based on their characteristics:
  + Cluster 0: Moderately Staffed
  + Cluster 1: Critically Understaffed
  + Cluster 2: Understaffed
  + Cluster 3: Well-Staffed

**Step 5: Visualization**

* A scatter plot was created to visualize the relationship between doctor shortfall percentage and PHC area covered, colored by cluster:
* plt.figure(figsize=(10, 6))
* colors = ['blue', 'green', 'orange', 'red']
* for i, cluster in enumerate(data['Cluster\_Name'].unique()):
* cluster\_data = data[data['Cluster\_Name'] == cluster]
* plt.scatter(cluster\_data['Doc\_Shortfall\_Pct'], cluster\_data['PHC\_Area\_Covered'],
* c=colors[i], label=cluster, s=100)
* for i, row in data.iterrows():
* if row['Doc\_Shortfall\_Pct'] > 30 or row['PHC\_Area\_Covered'] > 200:
* plt.text(row['Doc\_Shortfall\_Pct'] + 1, row['PHC\_Area\_Covered'], row['State'], fontsize=9)
* plt.title('Doctor Shortfall vs. PHC Area Covered by Segment')
* plt.xlabel('Doctor Shortfall (%)')
* plt.ylabel('PHC Area Covered (Sq. Km.)')
* plt.legend(bbox\_to\_anchor=(1.05, 1), loc='upper left')
* plt.tight\_layout()
* plt.savefig('segment\_scatter\_plot.png')
* plt.show()
  + **Purpose**: Highlighted cluster differences and identified outliers (e.g., states with high shortfalls or large coverage areas).
  + **Output**: A PNG file (segment\_scatter\_plot.png) was saved for reference.

**Step 6: Analysis of Imbalances and Disparities**

* **Staffing Imbalances**:
  + Identified states with disproportionate staffing ratios:
  + imbalanced = data[
  + ((data['Doc\_Staffing\_Ratio'] < 80) & (data['Nurse\_Staffing\_Ratio'] > 120)) |
  + ((data['Surgeon\_Staffing\_Ratio'] < 20) & (data['Nurse\_Staffing\_Ratio'] > 100))
  + ]
  + print('States with Imbalanced Staffing:')
  + print(imbalanced[['State', 'Doc\_Staffing\_Ratio', 'Nurse\_Staffing\_Ratio', 'Surgeon\_Staffing\_Ratio']].to\_string(index=False))
    - **Criteria**: Low doctor staffing (<80) with high nurse staffing (>120), or low surgeon staffing (<20) with high nurse staffing (>100).
    - **Output**: 18 states/UTs, including Andhra Pradesh, Kerala, and Chandigarh, showed imbalances.
* **Geographic Disparities**:
  + Identified states with high PHC area coverage:
  + high\_coverage = data[data['PHC\_Area\_Covered'] > data['PHC\_Area\_Covered'].quantile(0.75)]
  + print('States with High PHC Area Coverage:')
  + print(high\_coverage[['State', 'PHC\_Area\_Covered', 'Doc\_Shortfall\_Pct']].to\_string(index=False))
    - **Criteria**: PHC area coverage above the 75th percentile.
    - **Output**: 9 states, including Mizoram (369.84 sq. km.) and Chhattisgarh (168.08 sq. km.).

**Step 7: Segment Details**

* Detailed statistics for each cluster were printed:
* for cluster in data['Cluster\_Name'].unique():
* print(f'Segment: {cluster}')
* segment\_data = data[data['Cluster\_Name'] == cluster][['State'] + features + ['PHC\_Area\_Covered']]
* print(segment\_data.to\_string(index=False, float\_format='%.1f'))
  + **Purpose**: Provided a granular view of each segment’s staffing metrics and geographic coverage.

**4. Key Findings**

**Segment Profiles**

1. **Moderately Staffed (18 States/UTs)**:
   * **States**: Andhra Pradesh, Kerala, Tamil Nadu, Chandigarh, Lakshadweep, etc.
   * **Metrics**:
     + Doctor Shortfall: 0–14.7% (e.g., Andhra Pradesh: 0%, Arunachal Pradesh: 14.7%).
     + Nurse Shortfall: 0–14.7%.
     + Surgeon Shortfall: 75–100% (e.g., Kerala: 99.6%, Chandigarh: 100%).
     + Staffing Ratios: Doctor (85.3–233 spi3), Nurse (85.3–470.6), Surgeon (0–25).
     + PHC Area Coverage: 1.5–260.6 sq. km.
   * **Insight**: Adequate doctor and nurse staffing but severe surgeon shortages. Suitable for targeted surgeon recruitment.
2. **Critically Understaffed (7 States)**:
   * **States**: Bihar, Uttar Pradesh, Odisha, Jharkhand, etc.
   * **Metrics**:
     + Doctor Shortfall: 0–39% (e.g., Uttar Pradesh: 39%, Bihar: 6%).
     + Nurse Shortfall: 29.6–61.3% (e.g., Bihar: 61.3%).
     + Surgeon Shortfall: 81.6–95.5%.
     + Staffing Ratios: Doctor (61.0–161.5), Nurse (38.7–70.4), Surgeon (4.5–18.4).
     + PHC Area Coverage: 48.4–260.8 sq. km.
   * **Insight**: Severe shortages across all roles, especially nurses. Urgent recruitment and retention strategies are needed.
3. **Understaffed (10 States)**:
   * **States**: Chhattisgarh, Mizoram, Jammu & Kashmir, Madhya Pradesh, etc.
   * **Metrics**:
     + Doctor Shortfall: 0–56.6% (e.g., Chhattisgarh: 56.6%, Mizoram: 1.8%).
     + Nurse Shortfall: 0–12.2%.
     + Surgeon Shortfall: 37.9–100%.
     + Staffing Ratios: Doctor (43.4–140.2), Nurse (87.8–176.7), Surgeon (0–62.1).
     + PHC Area Coverage: 4.8–369.8 sq. km.
   * **Insight**: Inconsistent staffing with high vacancies in some states and large geographic coverage, requiring localized solutions.
4. **Well-Staffed (1 State)**:
   * **State**: Delhi.
   * **Metrics**:
     + Doctor Shortfall: 0%.
     + Nurse Shortfall: 0%.
     + Surgeon Shortfall: 0%.
     + Staffing Ratios: Doctor (420.0), Nurse (140.0), Surgeon (0 or missing).
     + PHC Area Coverage: 73.9 sq. km.
   * **Insight**: Exceptional staffing, likely due to urban advantages. A benchmark for best practices.

**Staffing Imbalances**

* **18 States/UTs** showed imbalances:
  + **Examples**:
    - Andhra Pradesh: Nurse (141.8), Surgeon (19.7).
    - Chandigarh: Nurse (470.6), Surgeon (0.0).
    - West Bengal: Nurse (275.5), Surgeon (0.0).
  + **Insight**: High nurse staffing contrasts with low surgeon or doctor staffing, indicating a need for specialized recruitment.

**Geographic Disparities**

* **9 States** with high PHC area coverage:
  + **Examples**:
    - Mizoram: 369.84 sq. km., Doctor Shortfall: 1.8%.
    - Jammu & Kashmir: 346.92 sq. km., Doctor Shortfall: 0%.
    - Chhattisgarh: 168.08 sq. km., Doctor Shortfall: 56.6%.
  + **Insight**: Large coverage areas exacerbate staffing challenges, particularly in rural states.

**Visualization Insights**

* The scatter plot highlighted:
* **Outliers**: Chhattisgarh (high shortfall, moderate coverage), Mizoram (low shortfall, high coverage).
* **Cluster Separation**: Moderately Staffed states cluster near low shortfalls, while Critically Understaffed and Understaffed states show higher shortfalls and variable coverage.
* **Insight**: Geographic and staffing challenges are interconnected, with rural states facing greater difficulties.

**5. Recommendations**

**Critically Understaffed States**

* **Action**: Launch mass recruitment drives for doctors and nurses in states like Bihar and Uttar Pradesh.
* **Incentives**: Offer housing, bonuses, or loan forgiveness to improve retention.
* **Training**: Expand medical training programs to address nurse shortages.

**Understaffed States with High Coverage**

* **Action**: Deploy mobile health units in states like Mizoram and Chhattisgarh to cover large areas.
* **Technology**: Implement telemedicine to extend doctor availability in remote regions.
* **Infrastructure**: Build additional PHCs to reduce coverage per center.

**Imbalanced Staffing**

* **Action**: Recruit surgeons in states like Kerala and Chandigarh to address critical shortages.
* **Reallocation**: Transfer excess nurses from overstaffed regions (e.g., Chandigarh) to understaffed ones.
* **Specialization**: Develop surgical training programs to build capacity.

**Moderately Staffed States**

* **Action**: Maintain current staffing levels while addressing surgeon shortages through partnerships with medical institutions.
* **Monitoring**: Use data analytics to prevent future imbalances.

**General Strategies**

* **Benchmarking**: Study Delhi’s urban staffing model for replicable practices.
* **Data-Driven Planning**: Invest in real-time workforce monitoring systems.
* **Policy**: Develop national guidelines for equitable surgeon distribution.

**6. Conclusion**

The segmentation analysis reveals stark disparities in healthcare staffing across Indian states and Union Territories. Critically Understaffed states like Bihar and Uttar Pradesh face severe shortages, while Understaffed states like Chhattisgarh and Mizoram struggle with large geographic coverage. Imbalanced staffing, particularly low surgeon ratios, is widespread. By tailoring interventions to each segment—ranging from mass recruitment to telemedicine and surgeon training—policymakers can address these challenges and improve healthcare access nationwide.