Wellness Centres Data Analysis Report

1. Dataset Description

1.1 Source

Public dataset from Data.gov.in – City-wise List of Wellness Centres. Contains structured data of wellness centres operating across various Indian cities under different categories such as AYUSH, Urban Wellness Units, and Community Health Centres.

1.2 Columns

- citycode – Unique identifier for each city

- wellnesscentercode – Unique code for each wellness centre

- category – Type of wellness centre (e.g., AYUSH, Urban, Rural)

- doctorcount – Number of doctors assigned to each centre

- latitude – Centre's latitude coordinate

- longitude – Centre's longitude coordinate

- wellnesscentrename – Official name of the wellness centre

- cityname – City where the centre is located

- wellnesscenternumber – Administrative number used for identification

- wellnesscenteraddress – Address or location of the centre

1.3 Data Overview

• Rows (Records): ~500 wellness centre entries

• Columns (Attributes): 10

• Data Type Mix: Categorical (City, Category, Centre Name) + Numerical (Doctor Count, Latitude, Longitude)

1.4 Data Quality

Few missing or invalid values observed in doctorcount column (e.g., non-numeric entries). Latitude and longitude values mostly valid for Indian regions. Overall, the dataset is clean, consistent, and suitable for PySpark-based analytics and visualization.

2. Operations Performed

2.1 Data Cleaning & Exploration

Loaded dataset into PySpark DataFrame (df_spark). Verified schema, data types, and null entries. Cleaned doctorcount column using regex to remove invalid (non-numeric) entries and converted it to integer. Checked unique values in category and cityname, and computed record and column counts.

2.2 Descriptive Analytics

• Category Distribution: Counted the number of wellness centres under each category.

• City-wise Distribution: Aggregated total centres per city.

• Doctor Availability: Summed doctor counts city-wise and category-wise.

• Missing Data Check: Identified centres lacking doctor data.

2.3 Visualization

Generated visualizations including:

- Pie Chart: Top 5 categories of wellness centres

- Bar Chart: City-wise distribution of centres

- Scatter Plot: Relationship between total centres and doctor availability

- Box Plot: Doctor count spread across categories

- Optional: Geographic map of centres using Folium

2.4 Advanced Analysis

Grouped and aggregated city-wise doctor statistics. Ranked cities by number of centres and doctors. Filtered centres with more than 5 doctors. Exported cleaned dataset and visual insights to CSV and PDF.

3. Key Insights

3.1 Category Insights

AYUSH Centres and Urban Wellness Units account for more than 60% of total wellness centres. Some rare categories exist with minimal records, showing specialized wellness facilities.

3.2 City-wise Distribution

Metro cities like Delhi, Mumbai, Bangalore, and Hyderabad have higher concentrations of wellness centres. Smaller cities show fewer centres, reflecting potential gaps in healthcare access.

3.3 Doctor Availability

Doctor count varies widely by city and category. Some centres have missing doctor data. Urban regions generally show higher average doctor availability per centre.

3.4 Data Quality Observations

Around 5–10% of doctorcount entries were invalid. After cleaning, numeric consistency improved. No major duplication or structural inconsistencies found.

3.5 Geographical Spread

Most centres are located in northern and southern India, with dense clusters in urbanized states. Central and north-eastern regions show sparse coverage.

4. Recommendations

4.1 Resource Allocation

Increase establishment of centres in underserved regions and balance doctor deployment between urban and rural areas.

4.2 Data Management

Standardize data entry fields and introduce validation checks to prevent invalid values and ensure accuracy.

4.3 Infrastructure Planning

Use city-level insights to plan future wellness expansion and infrastructure investment.

4.4 Analytical Extension

Combine with population data for better context and create predictive models for future resource needs.