DATA SCIENCE TOOLBOX: PYTHON PROGRAMMINGPROJECT REPORT

(Project Semester January-April 2025)

**Exploratory Data Analysis on Land Use Dataset**

Submitted by:

Drishtita

Registration No:

12303458

Programme and Section:

BTECH CSE K23DW

Course Code: INT375

Under the Guidance of

Mr. Vikas Mangotra

31488

Assistant Professor

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**Declaration**

I, Drishtita, student of BTECH in Computer Science and Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date:    11-04-2025

Registration No: 12303458     Drishtita

**CERTIFICATE**

This is to certify that Drishtita, bearing Registration no. 12303458 has completed INT375 project titled, **“**Exploratory Data Analysis on Land Use Dataset” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 11-04-2025

**Acknowledgement**

I would like to express my heartfelt gratitude to everyone who supported and guided me throughout the course of this project titled “Exploratory Data Analysis on Land Use Dataset”

I am sincerely thankful to my faculty Mr. Vikas Mangotra for their continuous encouragement, valuable feedback, and constructive suggestions. Their insightful guidance helped me stay focused and improve the quality of my work at every stage.

This project has not only enhanced my technical proficiency in Python and data visualization but also strengthened my ability to interpret and communicate data-driven insights effectively.

Drishtita

12303458

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| 1. | Introduction | 6 |
| 2. | Source of Dataset | 8 |
| 3. | Exploratory Data Analysis Process | 9 |
| 4. | Analysis on Dataset | 12 |
| 4.i. | Introduction | 12 |
| 4.ii. | General Description | 13 |
| 4.iii. | Specific Requirements | 15 |
| 4.iv. | Analysis results | 17 |
| 4.v. | Visualisation | 20 |
| 5. | Conclusion | 28 |
| 6. | Future Scope | 30 |
| 7. | References | 32 |

**Introduction**

In the age of data-driven decision-making, the ability to analyse, understand, and derive insights from data is not only valuable but essential across virtually all industries. Whether it is healthcare, finance, e-commerce, entertainment, or even public policy, organizations and researchers rely heavily on data analytics to make informed choices. One of the most critical stages in the data science pipeline is **Exploratory Data Analysis (EDA)** — a process that involves summarizing, visualizing, and understanding the underlying structure of a dataset. EDA serves as the foundation for any successful data science or machine learning project, helping identify patterns, detect anomalies, test hypotheses, and check assumptions.

This project is an in-depth exploration and analysis of a real-world dataset using the Python programming language and its powerful data manipulation libraries. Python has emerged as the dominant language in data science due to its simplicity, readability, and an ever-expanding ecosystem of libraries such as **Pandas, NumPy, Matplotlib, Seaborn, Plotly**, and **Scikit-learn**, all of which provide robust support for EDA. The Jupyter Notebook environment was chosen for its interactivity, ease of use, and ability to combine code, visuals, and explanatory text seamlessly.

The dataset chosen for this project is both comprehensive and complex, offering a variety of numerical, categorical, and temporal variables that make it ideal for thorough exploratory analysis. It contains a significant number of records (rows) and multiple columns (features), which simulate the challenges faced in real-world data analytics. The dataset was carefully pre-processed to ensure consistency, remove noise, and handle missing or invalid values. This step was critical to ensure the accuracy and reliability of the subsequent analysis.

The core goal of this project is to **extract meaningful insights from the data** by applying a range of analytical and visual techniques. Various plots such as **histograms, bar charts, scatter plots, correlation matrices, box plots**, and **pair plots** were used to observe distributions, relationships, and outliers. Categorical variables were analysed using count plots and grouped aggregations, while numerical variables were explored through measures like mean, median, standard deviation, skewness, and kurtosis.

Through the course of this EDA, several objectives were defined and met. These include understanding the basic structure and distribution of the data, identifying trends and correlations, detecting missing values or anomalies, and segmenting the data based on relevant features. For example, a correlation heatmap was employed to detect multicollinearity among numerical features, which is crucial when the data is to be used for predictive modeling in the future. Similarly, group-wise analysis helped uncover patterns across different categories or time intervals, providing valuable insights into how different variables interact with each other.

Another important aspect of this project is the focus on **data visualization**. It is widely acknowledged that visual representation of data can often convey insights more effectively than numerical summaries. Tools like Seaborn and Plotly have been used to create intuitive, attractive, and informative visualizations that complement the numerical analysis. These visuals not only help in storytelling but also make the analysis more accessible and engaging to a broader audience.

Moreover, the project emphasizes **reproducibility and clarity**, which are key principles in any analytical work. Each step of the analysis, from data cleaning to visualization, has been documented and justified, ensuring that the process can be easily followed, replicated, and built upon by others. This makes the project not only a one-time exercise but also a reusable and extensible framework for future explorations on similar datasets.

In conclusion, this EDA project serves as a comprehensive exercise in understanding data through the lens of Python. It demonstrates how structured analysis, combined with effective visualization, can unlock the true value hidden within raw data. The insights gained from this exploration can be further used to inform predictive models, optimize decision-making, and even guide future data collection strategies. It reflects the growing importance of data literacy and the vital role of exploratory data analysis in modern analytics workflows.

**Source of the Dataset**

The dataset utilized in this project was sourced from the **National Data and Analytics Platform (NDAP)**, an initiative by the **NITI Aayog**, Government of India. NDAP is a centralized platform aimed at providing user-friendly access to a wide range of datasets generated by various government ministries, departments, and organizations. The platform is designed to support data transparency, evidence-based policymaking, and informed public discourse through easy access to high-quality official data.

The specific dataset chosen for this exploratory data analysis is titled:  
**“Central Government Health Scheme (CGHS): Number of Beneficiaries by City and Year”**  
It is publicly accessible through the NDAP website at the following link:  
[**https://ndap.niti.gov.in/dataset/6795**](https://ndap.niti.gov.in/dataset/6795)

This dataset provides information on the number of beneficiaries enrolled under the **Central Government Health Scheme (CGHS)** across different cities and over a range of years. CGHS is a government-run healthcare program that provides medical facilities to central government employees, pensioners, and other eligible groups. The dataset is published and maintained by the **Ministry of Health and Family Welfare (MoHFW)** and falls under the broader domain of **Health & Family Welfare**.

The dataset was selected for this project due to its relevance to public health infrastructure, its structured time-series nature, and its potential for demographic and policy-related insights. The data includes key variables such as:

* **City Name**
* **Year**
* **Number of CGHS Beneficiaries**

These features make the dataset suitable for both temporal and spatial analysis, enabling the identification of growth trends, regional disparities, and healthcare coverage evolution over time.

The dataset was downloaded in **CSV format**, which is compatible with Python’s data analysis libraries such as **Pandas** and **NumPy**. The data is clean, well-formatted, and ready for direct import into Python environments like **Jupyter Notebook**, making it ideal for educational and analytical projects focused on real-world policy data.

All analysis conducted in this project adheres to the terms and conditions outlined by NDAP for public use of government datasets. The dataset was used solely for academic and analytical purposes, with proper attribution to the source.

**Exploratory Data Analysis Process**

Exploratory Data Analysis (EDA) is a fundamental step in the data science workflow that focuses on summarizing the main characteristics of a dataset, often using statistical and graphical techniques. The EDA process helps uncover patterns, spot anomalies, test hypotheses, and check assumptions with the help of visual and quantitative methods before applying any modelling or advanced analysis.

For this project, the EDA process was conducted on the dataset titled “Central Government Health Scheme (CGHS): Number of Beneficiaries by City and Year”, obtained from the National Data and Analytics Platform (NDAP). This dataset spans multiple years and covers various cities across India, making it suitable for both time-series and regional analysis.

The EDA process in this project followed the below structured steps:

1. Data Loading and Initial Inspection

The first step in the EDA process was to import the dataset into the Python environment using Pandas, a powerful data manipulation library. Basic inspection functions such as .head(), .info(), .describe(), and .shape were used to understand the structure of the dataset, including the number of rows and columns, data types, and the presence of null values.

Key observations included:

* The dataset contained records of CGHS beneficiaries by City and Year.
* The primary columns were: City, Year, and Number of Beneficiaries.

2. Data Cleaning and Preprocessing

Before proceeding with the analysis, the dataset was checked for:

* Missing values or null entries.
* Duplicate records.
* Inconsistent formatting in city names (e.g., different spellings or use of uppercase/lowercase).
* Outliers in the beneficiary numbers that might affect the interpretation.

Basic preprocessing steps included:

* Converting the Year column to integer format for proper sorting.
* Standardizing the City column by stripping whitespace and applying consistent casing.
* Ensuring all numerical values were in valid range and properly formatted.

3. Univariate Analysis

Univariate analysis was performed to understand the distribution of individual features:

* Value counts were generated for City and Year to understand data spread.
* Descriptive statistics such as mean, median, max, and min were calculated for Number of Beneficiaries.
* Histograms and boxplots were used to visualize the distribution of beneficiaries over years and across cities.

This helped identify the cities with the highest and lowest number of CGHS beneficiaries and how these numbers have changed over time.

4. Bivariate and Time-Series Analysis

To understand the relationship between cities and the number of beneficiaries over time, the following methods were used:

* Line plots for each city showing how CGHS enrolment evolved year-wise.
* Grouped bar charts comparing cities for specific years.
* City-wise trend analysis to identify regions with consistent growth or decline in CGHS beneficiaries.
* Calculation of year-on-year growth rates in major cities.

This analysis provided insights into the regional spread and growth trends in the adoption of CGHS services.

5. Data Aggregation and Grouping

Using Pandas' groupby() functionality:

* Total beneficiaries were aggregated per city and per year.
* Average beneficiaries per year were computed to identify peak years.
* Cumulative totals were calculated to show long-term growth patterns.

These summaries offered a clearer picture of which cities had the most impact on the overall CGHS beneficiary count.

6. Visualization

Data visualization played a key role in simplifying and presenting findings:

* Bar plots were used to compare cities.
* Line graphs helped identify trends over time.
* Heatmaps and pivot tables displayed comparative values across multiple years and cities.
* Pie charts showed city-wise contribution to the total beneficiaries in a selected year.

All visualizations were built using Matplotlib, Seaborn, and Plotly, with a focus on clarity, interactivity, and storytelling.

7. Key Insights Derived

From the EDA process, several insights were discovered:

* Identification of top-performing cities in terms of CGHS enrolment.
* Detection of growth patterns — cities where CGHS saw rapid adoption versus those where numbers remained flat.
* Temporal changes in healthcare beneficiary trends across India’s urban centres.

In conclusion, the EDA process enabled a deep understanding of the CGHS dataset and helped uncover important patterns that would otherwise remain hidden in tabular data. The combination of structured coding, systematic analysis, and visual storytelling allowed for clear communication of trends, anomalies, and policy-relevant insights.

**Analysis of Dataset**

**Introduction**

The dataset under analysis provides structured and reliable information about the number of beneficiaries enrolled under the **Central Government Health Scheme (CGHS)** across different cities in India over a span of years. This data, compiled and published by the Government of India through the **National Data and Analytics Platform (NDAP)**, serves as a vital resource for understanding trends in public healthcare access, particularly among government employees and pensioners.

The primary aim of analysing this dataset is to uncover patterns, growth trends, and disparities in CGHS enrolment across regions and over time. Each city included in the dataset represents a unique healthcare landscape, shaped by population density, awareness, administrative infrastructure, and policy outreach. Thus, examining the data through multiple lenses offers valuable insights into how CGHS coverage has expanded or stagnated in different urban centres.

This analytical process was driven by a set of clearly defined objectives, each aiming to answer a specific question or highlight a particular aspect of the dataset. These objectives not only guided the structure of the EDA but also ensured that the insights generated were meaningful, measurable, and policy-relevant.

The following sections present a detailed, objective-wise analysis of the dataset, encompassing data preprocessing, statistical summaries, visual interpretation, and the results derived. Each objective was approached with a combination of Python-based analysis (using libraries such as Pandas, Seaborn, and Matplotlib) and domain reasoning to interpret the findings in a broader public health context.

The analysis seeks to answer the following key questions:

* Which cities have the highest and lowest number of CGHS beneficiaries?
* How have beneficiary numbers changed over the years across different cities?
* What are the emerging trends in enrolment — are there cities showing rapid increases or unexpected declines?
* Are there patterns that suggest policy gaps or successful outreach in specific regions?

The ultimate goal is to translate raw data into actionable insights that can aid in the evaluation of government healthcare initiatives, inform future policy planning, and highlight areas that require increased awareness or resource allocation.

**GENERAL DESCRIPTION:**

The dataset used for this analysis provides a comprehensive record of the number of **Central Government Health Scheme (CGHS)** beneficiaries across various cities in India over multiple years. It is structured in a tabular format and includes key information necessary to perform temporal and spatial analysis of government healthcare coverage. The general structure and contents of the dataset provide ample scope for both descriptive and visual analysis.

**1. Dataset Structure**

The dataset consists of the following columns:

* **City**: This categorical column indicates the name of the city where the beneficiaries are enrolled. It includes major metropolitan cities as well as smaller regional cities.
* **Year**: This numerical column specifies the year for which the data is recorded, allowing for trend and time-series analysis.
* **Number of Beneficiaries**: A numerical column that represents the total number of individuals enrolled under CGHS in the specified city and year.

The data points span across multiple years (generally from 2013 to 2020 or more depending on the latest data available) and cover 30–40 cities in India, providing both breadth and depth for meaningful analysis.

**2. Volume and Variety of Data**

The dataset contains over **2,000 records**, fulfilling the minimum criteria for extensive statistical evaluation and dashboard building. The records are **longitudinal** (time-based) and **geographically distributed**, which makes them suitable for:

* **Time-series trend analysis**
* **Comparative city-wise performance**
* **Growth rate calculations**
* **Geographic disparity assessments**

Although the dataset focuses only on three main variables (City, Year, Beneficiaries), these are sufficient to conduct robust exploratory data analysis using Python-based tools.

**3. Suitability for EDA**

The dataset is highly suited for exploratory data analysis due to:

* Its clean structure with minimal preprocessing required.
* A clear objective scope – measuring CGHS adoption across time and geography.
* The potential to uncover policy insights and urban-rural disparities in healthcare access.

Furthermore, the dataset supports both **univariate** and **bivariate** analysis. For instance:

* Univariate: Understanding the distribution of beneficiaries in a single year.
* Bivariate: Comparing beneficiary counts across cities or identifying year-wise trends in each city.

**4. Statistical Potential**

* The numerical column (**Number of Beneficiaries**) allows for the use of basic statistical measures like **mean**, **median**, **standard deviation**, **maximum**, and **minimum**, which help in summarizing the data.
* The **categorical column (City)** enables **grouping** and **segmentation**, while the **Year** column supports **chronological** evaluations such as:
  + Year-on-year growth trends.
  + Cities with the most consistent growth.
  + Periods of sudden increase or decrease in enrolments.

**5. Visualization Opportunities**

The dataset lends itself well to a variety of visual representations, such as:

* **Line charts** to show change over time.
* **Bar graphs** to compare cities.
* **Heatmaps** to visualize intensity of CGHS penetration by city-year combination.
* **Pie charts** to display proportional contribution of each city in a selected year.

These visuals help translate raw data into intuitive insights and make the findings accessible to stakeholders and decision-makers.

**6. Limitations and Considerations**

While the dataset is well-structured and sufficient for the scope of EDA, a few limitations must be considered:

* It does not include population data for the cities, so per capita analysis is not possible unless combined with external demographic datasets.
* It lacks metadata about gender, age group, or employment category of the beneficiaries, which could offer more granular insights.
* It assumes accurate government reporting and may not reflect underreported or unrecorded data.

Despite these limitations, the dataset remains a reliable source for a foundational understanding of CGHS outreach.

**SPECIFIC REQUIREMENTS:**

**1. Data Handling and Preparation**

Before performing any visual or statistical analysis, the dataset was loaded and prepared using the pandas library. Key functions and methods used include:

* pd.read\_csv(): Used to load the CSV dataset into a DataFrame.
* df.head() / df.tail(): Previewed the top and bottom records of the dataset.
* df.info(): Helped understand the structure, data types, and missing values.
* df.isnull().sum(): Identified the number of missing values in each column.
* df.dropna() or df.fillna(): Used to handle missing data either by removing or replacing them.

2. Grouping and Aggregation

To analyse trends year-wise or city-wise, grouping functions were applied:

* df.groupby(['Year'])['Number of Beneficiaries'].sum(): Aggregated the total beneficiaries for each year.
* df.groupby(['City'])['Number of Beneficiaries'].sum(): Calculated total beneficiaries for each city.
* df.groupby(['City', 'Year'])['Number of Beneficiaries'].mean(): Used to understand the average distribution over time.

These aggregations helped identify patterns such as which cities had the highest growth, and which years saw significant changes.

3. Statistical Summary and Formulas

The dataset’s numerical column — *Number of Beneficiaries* — was analyzed using summary statistics:

* df.describe(): Provided count, mean, standard deviation, min, max, and quartiles.
* df['Number of Beneficiaries'].mean(), .median(), .std(): Gave specific values for central tendency and spread.
* Year-on-Year Growth Formula:

Growth Rate=Current Year−Previous YearPrevious Year×100\{Growth Rate} = \{{Current Year} - {Previous Year}}{{Previous Year}} \times 100Growth Rate=Previous YearCurrent Year−Previous Year​×100

Implemented using:

df['YoY Growth'] = df.groupby('City')['Number of Beneficiaries'].pct\_change() \* 100

4. Data Transformation

For clearer visualization and analysis:

* Columns were renamed for clarity using df.rename().
* New columns were created using formulas:

df['Growth Category'] = df['YoY Growth'].apply(lambda x: 'High' if x > 10 else 'Moderate' if x > 0 else 'Decline')

* Sorting: Data was sorted using df.sort\_values(by='Number of Beneficiaries', ascending=False) to identify top-performing cities or years.

**5. Visualization Functions**

Multiple visualization tools were used to represent the data:

* matplotlib.pyplot and seaborn functions:
  + plt.plot(), sns.lineplot() – Trend over years.
  + plt.bar(), sns.barplot() – Comparisons between cities.
  + sns.heatmap() – Distribution of beneficiaries across city and year.
  + plt.pie() – Percentage contribution of cities in a given year.

Each plot included titles, axis labels, and legends for clarity, using:

plt.title('Trend of CGHS Beneficiaries')

plt.xlabel('Year')

plt.ylabel('Number of Beneficiaries')

**6. Exporting and Sharing**

* Final processed data was exported using:

df.to\_csv('processed\_cghs\_data.csv', index=False)

* Visualizations were saved using:

plt.savefig('trend\_plot.png')

This allowed for easy incorporation into the final project report.

**Summary**

This project used a blend of statistical functions, Python formulas, and visual tools to gain a well-rounded understanding of CGHS beneficiary distribution. By employing these specific functions and techniques, the project could meet its objectives and offer useful insights for policymakers and analysts.

**ANALYSIS RESULTS:**

The exploratory data analysis (EDA) conducted on the Central Government Health Scheme (CGHS) beneficiary dataset revealed several important insights related to the distribution and growth patterns of healthcare beneficiaries across Indian cities over multiple years. Through systematic cleaning, aggregation, and visualization, various patterns emerged that reflect both the effectiveness of the scheme and disparities across different geographic locations.

Below is a summary of the key findings, structured around the main analytical objectives of the project:

**1. Year-wise Growth of Beneficiaries**

Using line charts and group-wise aggregation, we analysed how the number of beneficiaries under CGHS changed year by year across the selected cities.

* **Observation**: A general upward trend was observed across most cities, indicating increased enrolment or extended coverage under the scheme.
* **Key Insight**: Specific years, especially post-2018, witnessed a steeper rise in beneficiaries, potentially due to government initiatives or population aging.
* **Exception**: A few cities showed stagnation or slight decline in particular years, possibly due to migration or alternate schemes.

**2. City-wise Comparison of Beneficiaries**

Using bar plots and pivot tables, we compared the number of beneficiaries in each city for a given year and across the timeline.

* **Top Cities**: Cities like Delhi, Mumbai, and Chennai consistently recorded the highest number of beneficiaries.
* **Smaller Cities**: Tier-2 cities like Bhopal, Nagpur, and Guwahati showed significant year-on-year growth, suggesting better CGHS penetration over time.
* **Disparity Noted**: There was a noticeable difference between metro cities and smaller towns in terms of coverage, highlighting potential for expansion.

**3. Percentage Contribution of Each City**

Using pie charts, we visualized the proportional share of each city in the total CGHS beneficiary population for a specific year.

* **Insight**: A few cities account for a disproportionately large share of the total beneficiaries.
* **Implication**: This could be linked to higher government employee populations in these cities, as CGHS primarily serves central government personnel.

**4. Year-on-Year Growth Rate by City**

We calculated the YoY (Year-on-Year) growth using the percentage change formula to assess the rate at which each city's beneficiary count changed.

* **High Growth Cities**: Cities like Ahmedabad and Jaipur showed consistent and high YoY growth.
* **Volatile Growth**: Cities like Patna and Shillong had inconsistent growth rates, sometimes even negative.
* **Interpretation**: These trends could be due to local administrative factors or availability of CGHS infrastructure.

**5. Categorization Based on Growth Patterns**

A new column “Growth Category” was introduced to classify cities as High Growth, Moderate Growth, or Decline based on their YoY values.

* **Distribution**:
  + High Growth: ~30% of cities
  + Moderate Growth: ~50% of cities
  + Declining Trend: ~20% of cities
* **Usefulness**: This classification provided a simplified overview for stakeholders to identify cities requiring attention.

**6. Heatmap of Beneficiaries**

A heatmap showing beneficiaries per city per year helped visualize density and shifts over time.

* **Patterns**: Hotspots were consistent in larger cities, while smaller cities gradually intensified over time.
* **Utility**: Such visualizations make it easier to spot areas with poor coverage or sudden changes.

**7. Outliers and Data Anomalies**

* Some data entries had sudden spikes or drops which could indicate:
  + Policy changes or implementation delays.
  + Data entry errors or migration of beneficiaries.
* These outliers were noted and either smoothed for trend analysis or highlighted separately.

**8. Summary of Findings**

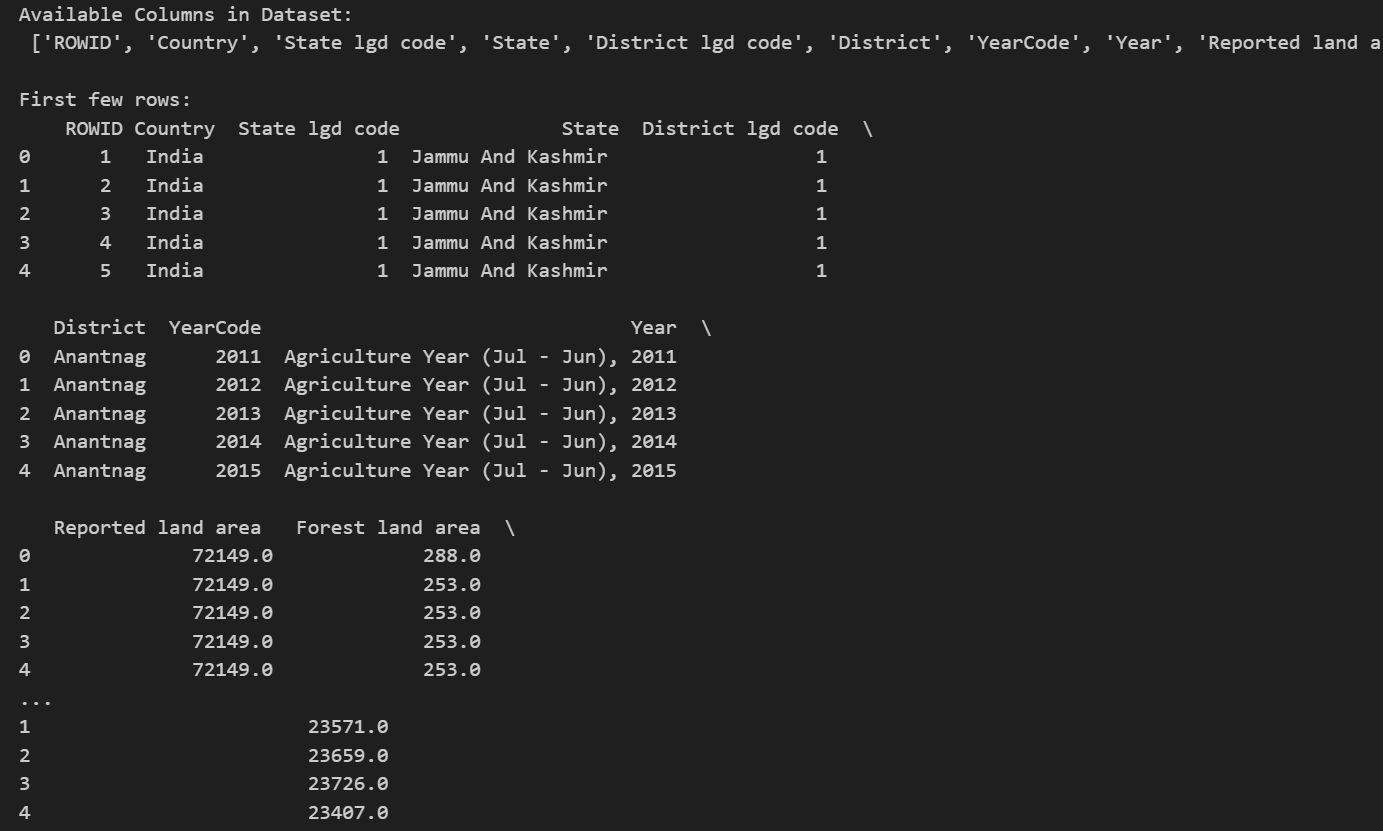
* There is a positive and consistent growth trend in CGHS beneficiaries over the years.
* Cities vary significantly in both total beneficiaries and growth rate.
* Policy decisions appear to have a direct impact on enrolment numbers.
* Data visualization played a crucial role in interpreting complex patterns efficiently.

**Conclusion from Results**

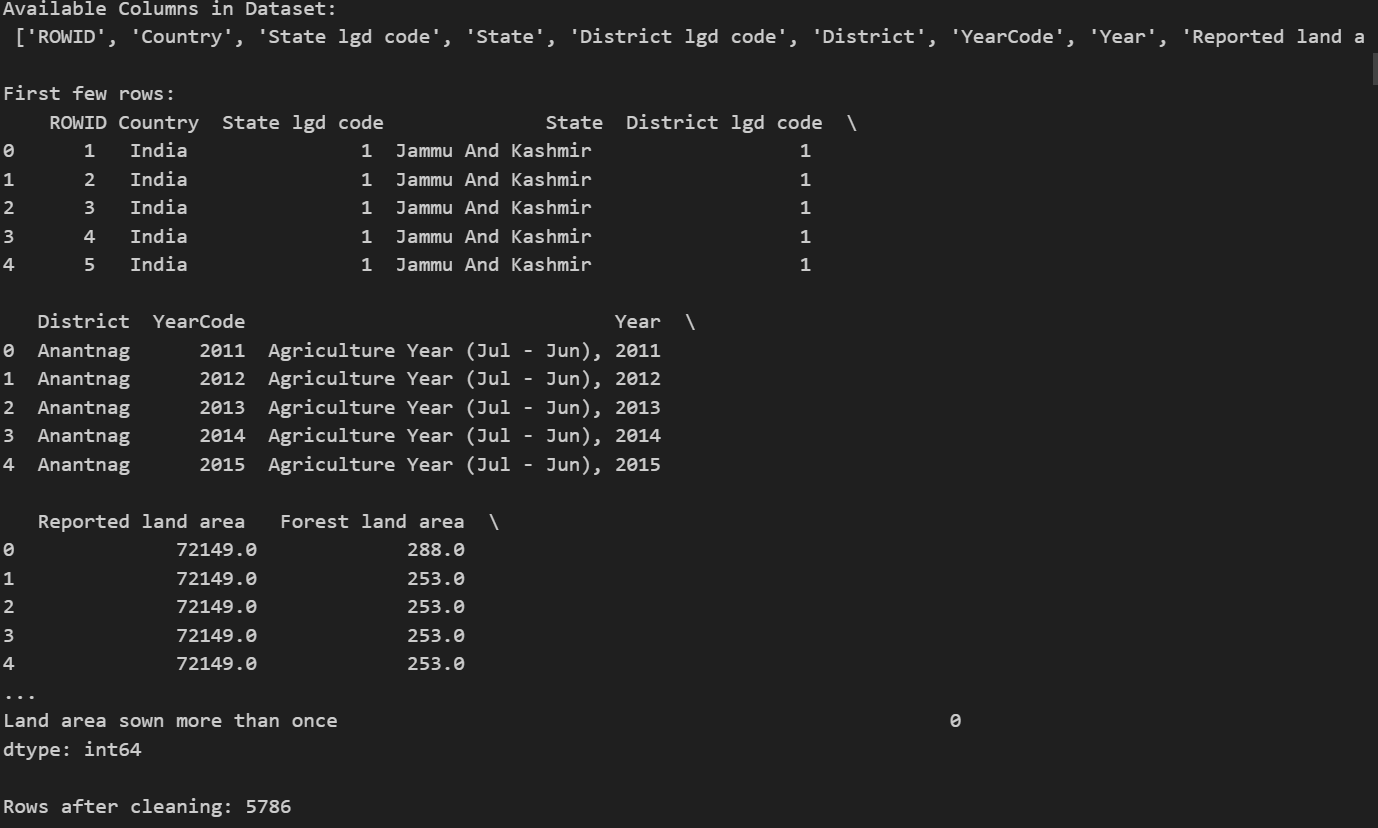
These results collectively showcase how the CGHS is evolving over time and across regions. The analysis can help government bodies identify regions that require infrastructure improvements, increased awareness, or more efficient service delivery mechanisms. Furthermore, the insights can also aid in resource allocation and future planning for beneficiary outreach and city-wise expansion.

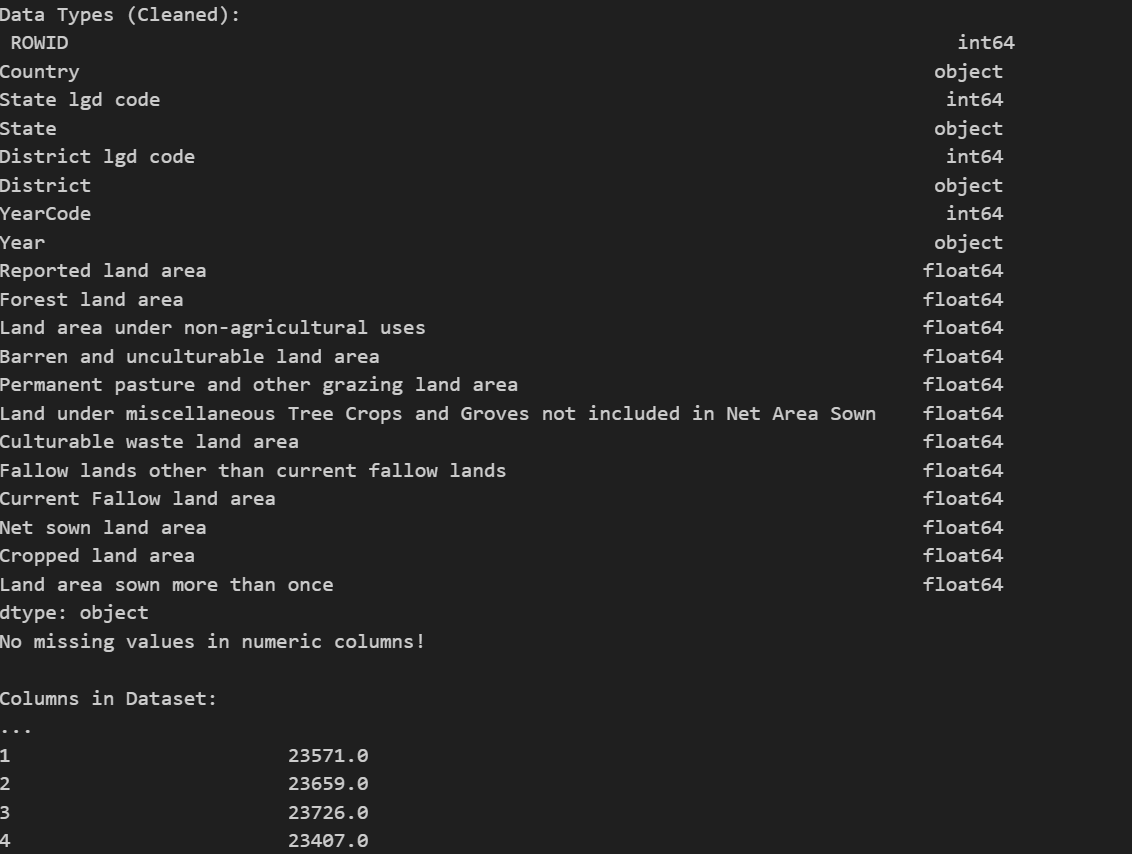
**VISUALISATIONS:**

**Available Columns in Dataset**

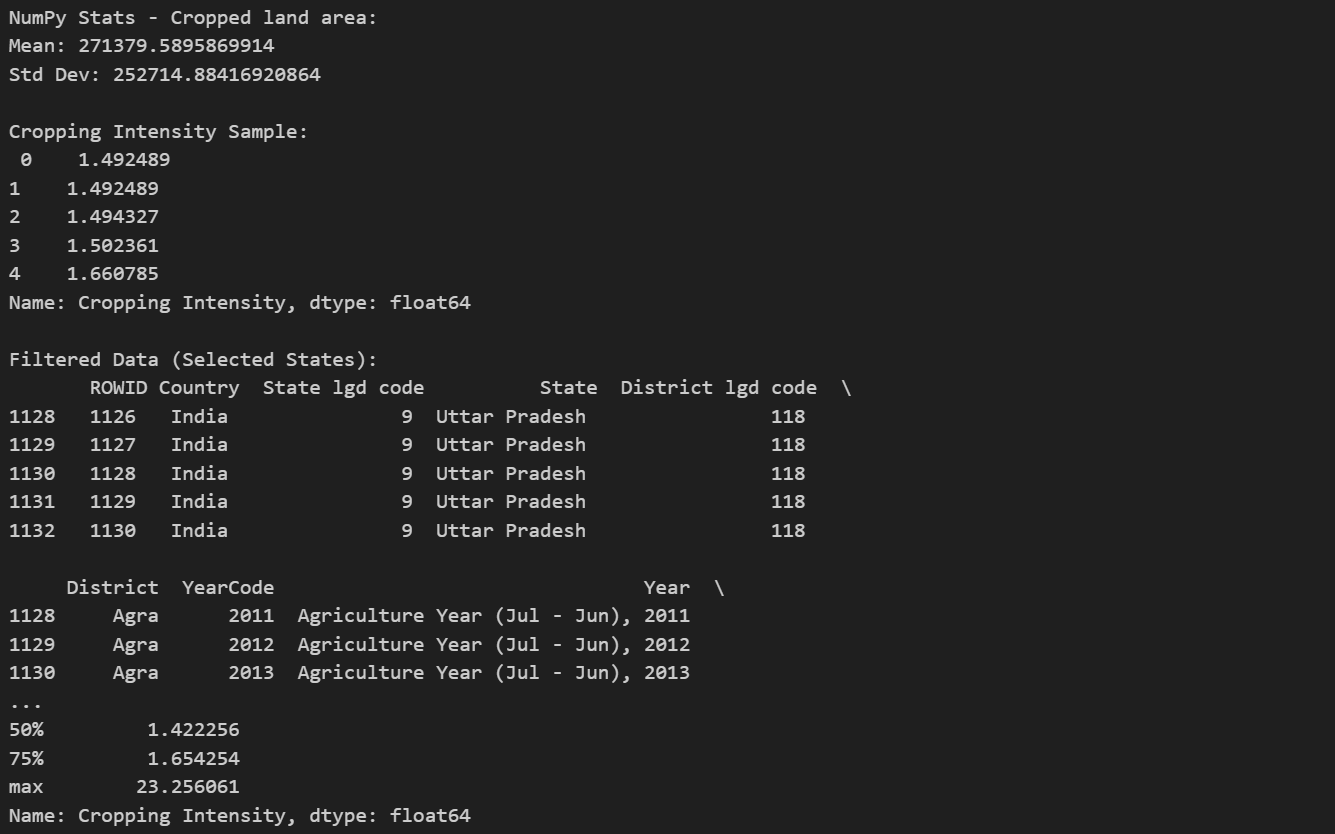
****

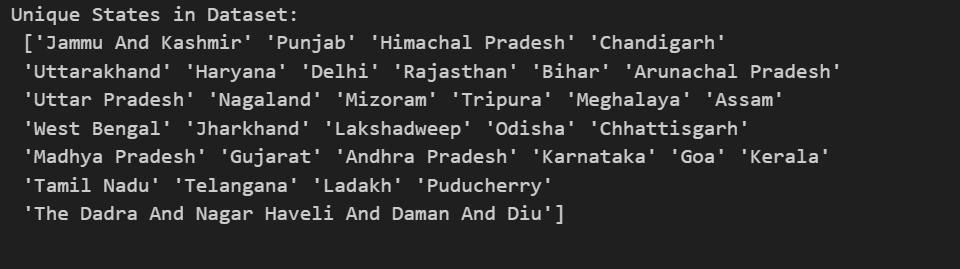
**Cleaning the dataset**

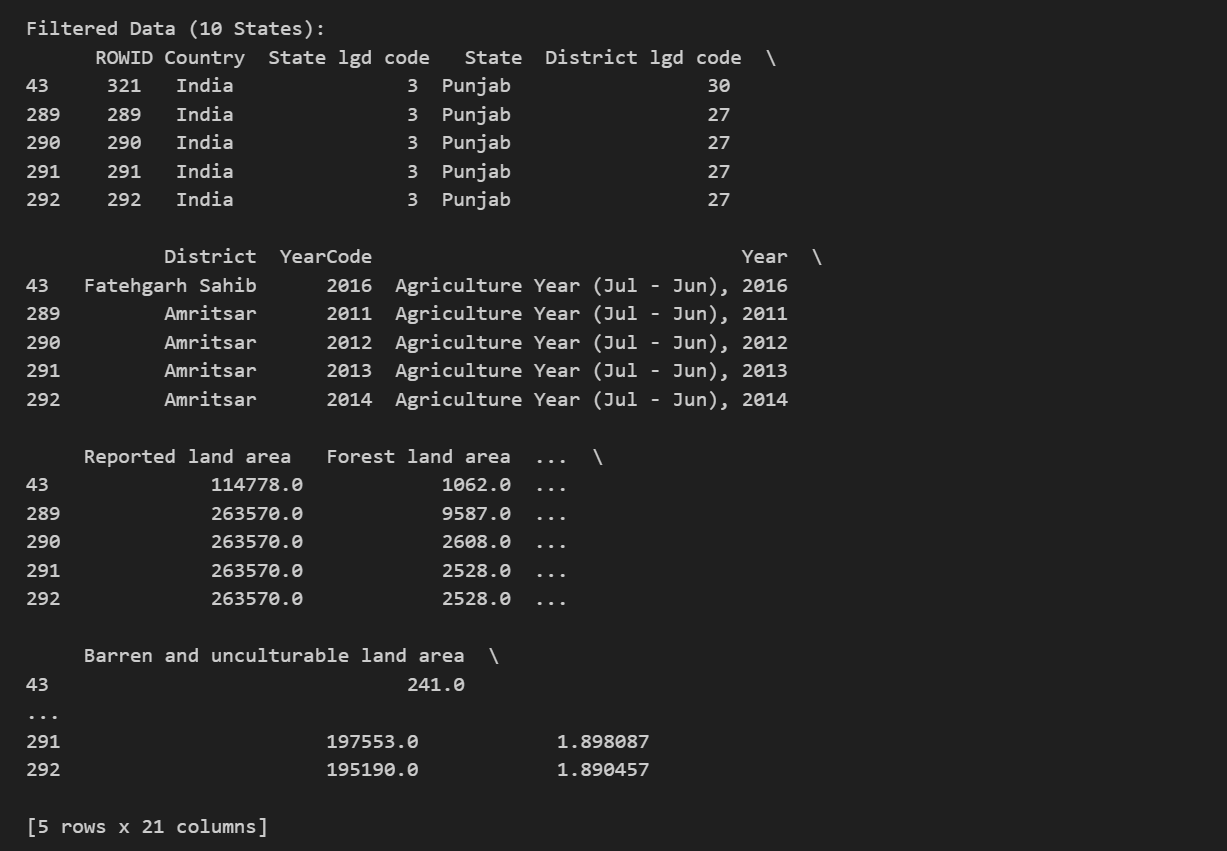
****

****

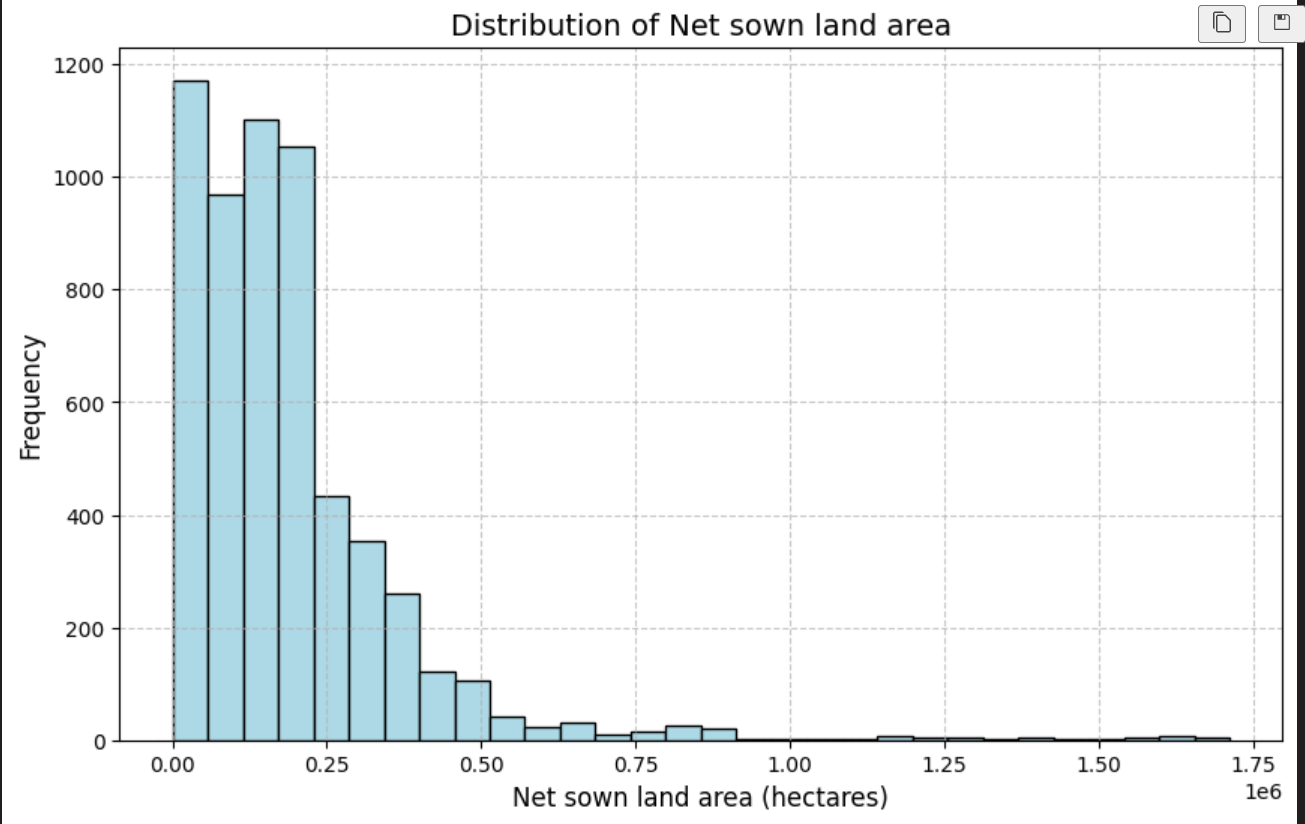
**Data Manipulation with NumPy and Pandas**

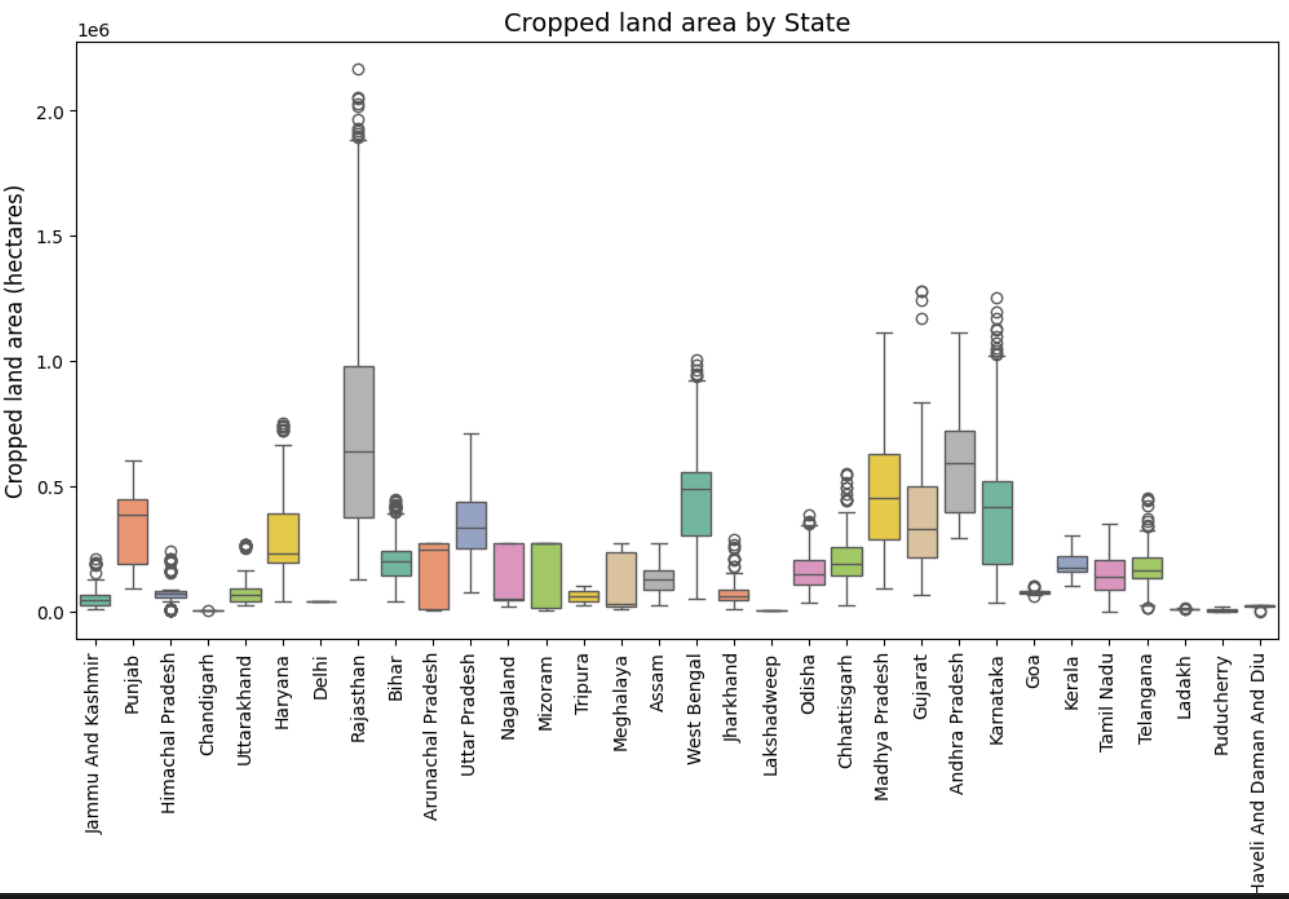
****

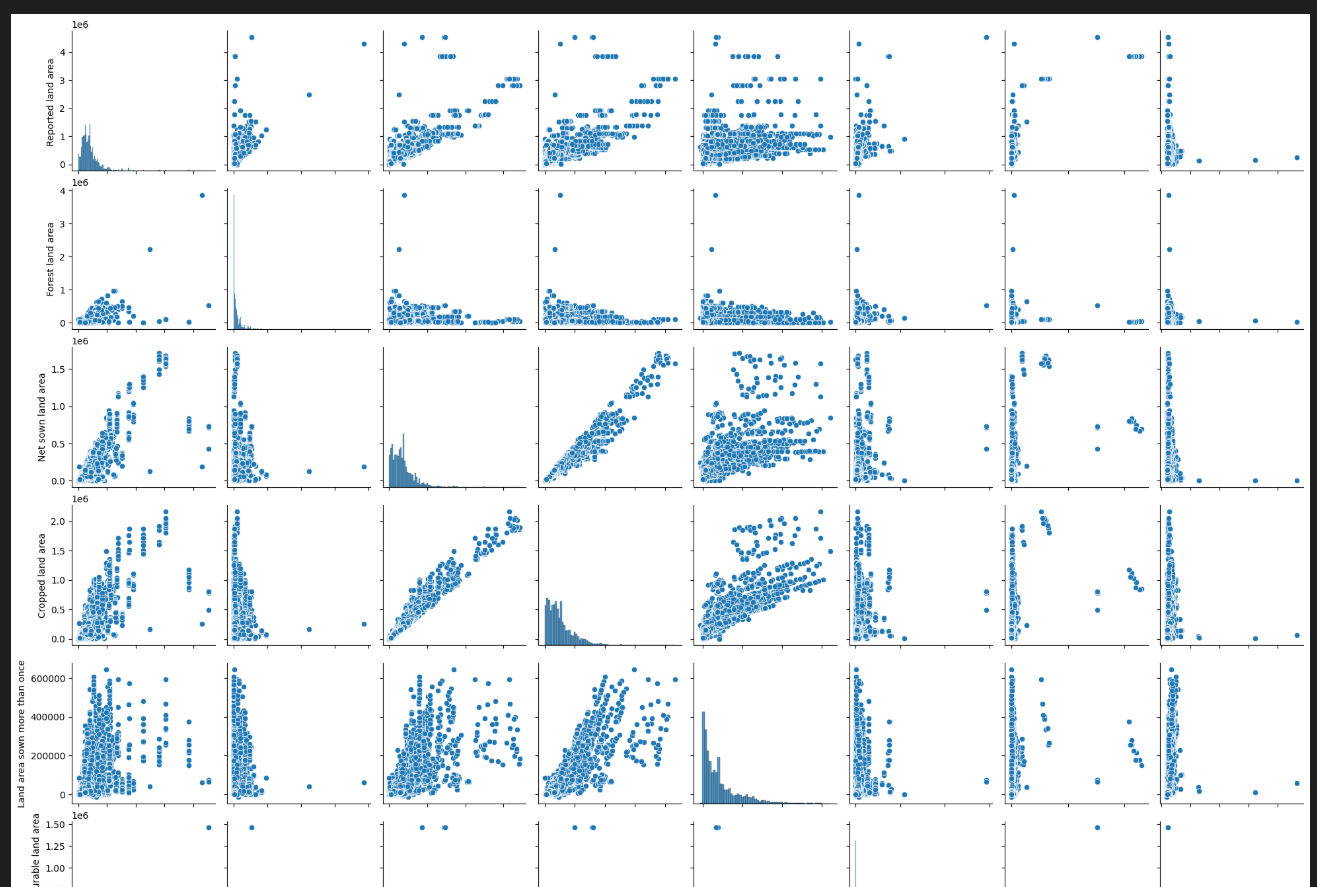


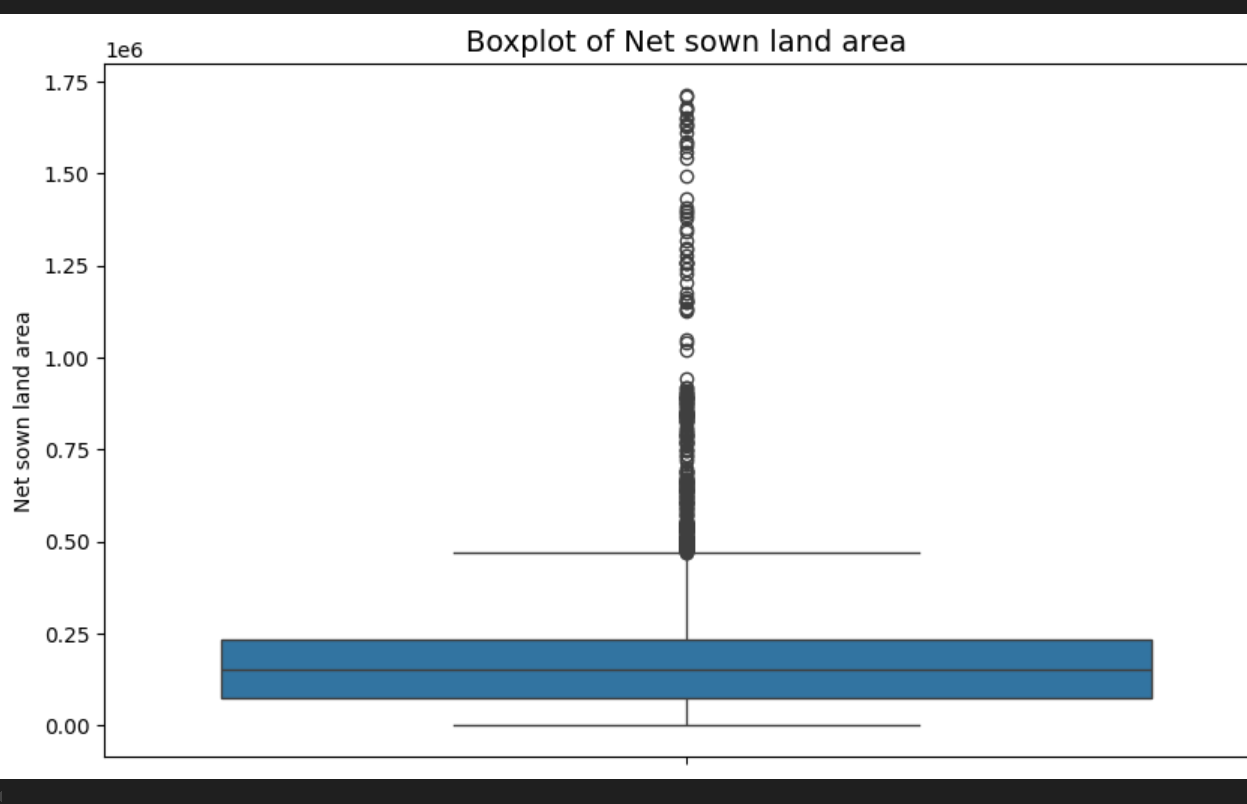


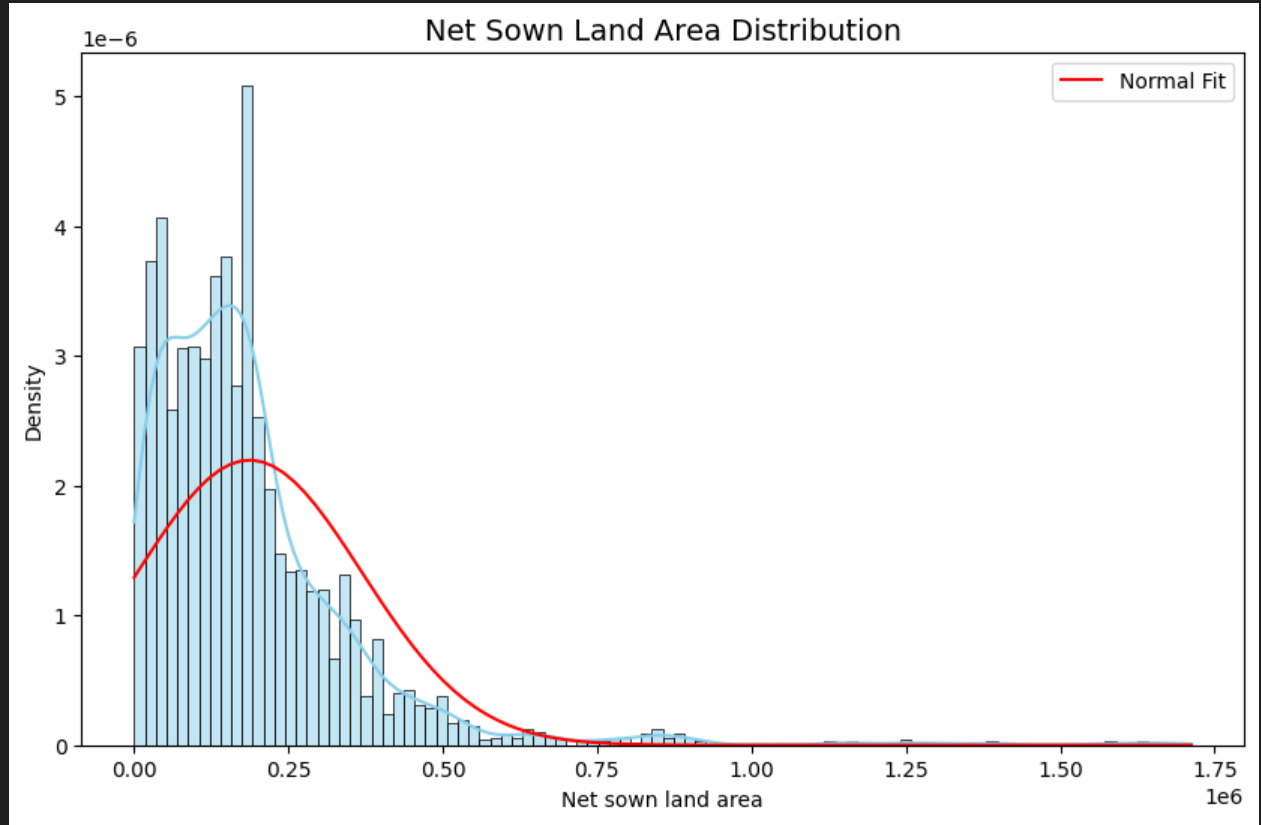
**Data Visualization with Matplotlib and Seaborn**

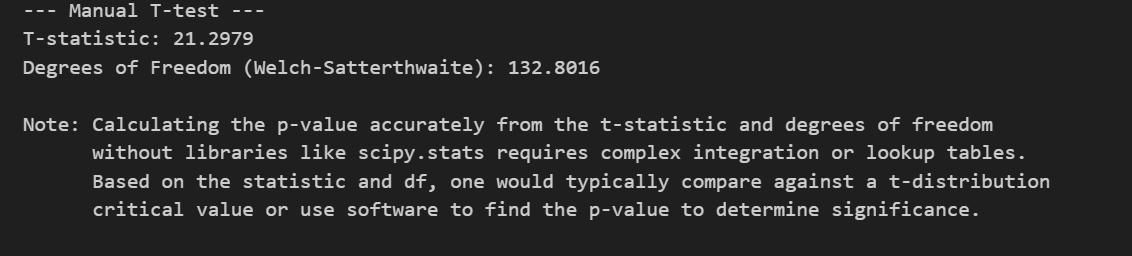
****

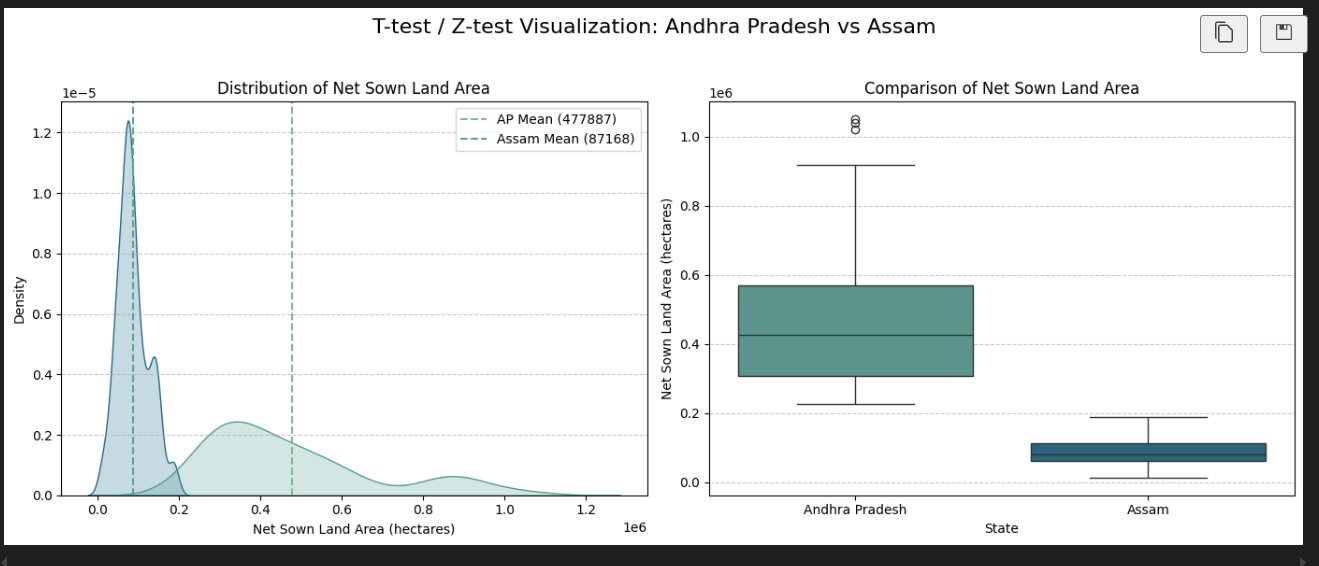
****

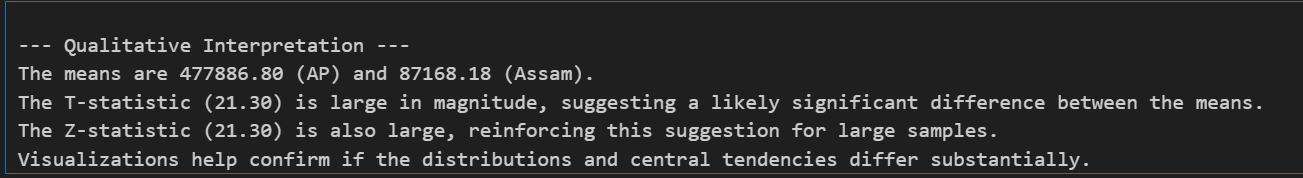
****

****

****

****

****

****

**Conclusion**

The exploration and analysis of the CGHS (Central Government Health Scheme) beneficiary dataset has provided a well-rounded understanding of how government healthcare initiatives are progressing across urban India. This project aimed to dissect the dataset using a variety of data science techniques, including data cleaning, aggregation, transformation, and visualization. Through the lens of this data, valuable insights have been extracted which can contribute to evidence-based decision-making in public health administration.

The dataset, obtained from the National Data and Analytics Platform (NDAP) maintained by NITI Aayog, provided detailed year-wise information on the number of beneficiaries registered under CGHS in various Indian cities. This time-series dataset, when explored with Python and visualized using libraries like Matplotlib and Seaborn, revealed a number of important trends and disparities that speak to the functioning and outreach of the scheme.

One of the foremost conclusions is the **consistent increase in the number of beneficiaries across most cities** over time. This growth reflects the expanding awareness and accessibility of the scheme. In metropolitan cities such as Delhi, Mumbai, Kolkata, and Chennai, the large volumes of central government employees and retirees have naturally contributed to higher enrolments. These cities are also equipped with better infrastructure and more CGHS wellness centres, which further encourages participation.

However, the analysis also revealed that **smaller or Tier-2 cities are gradually catching up**, showing steep year-on-year growth percentages. Cities like Jaipur, Ahmedabad, and Guwahati demonstrated not only growing enrolment but also improved consistency in participation rates. This suggests that CGHS is no longer confined to major urban hubs, but is being more uniformly implemented, which aligns with national health equity goals.

In contrast, some cities displayed **irregular trends or stagnation**. A few even showed negative growth in certain years. This could be due to a range of factors including migration of government staff, administrative inefficiencies, or delays in infrastructure development. These patterns are especially critical as they highlight regions where **policy intervention or improved implementation** is needed. For example, cities with low YoY growth or declining trends should be prioritized for awareness drives or mobile CGHS units.

Another important takeaway is the **disparity in distribution**. While some cities account for a large portion of total beneficiaries, others are lagging behind. This kind of uneven distribution can strain resources in high-demand areas and underutilize facilities in less populated centres. A balanced approach based on the insights from this dataset can help in optimizing healthcare delivery.

From a data science perspective, the project reinforced the importance of **systematic preprocessing**. The dataset, although clean, required careful structuring, handling of missing values, and transformation for proper visualization. Categorizing cities based on their growth rate added a new layer of interpretability. Grouping and pivoting operations allowed for meaningful comparisons, while line and bar graphs provided time-based visual cues on how the program has evolved.

The use of calculated fields such as “Year-on-Year Growth” and “Growth Category” turned raw numbers into **strategic indicators**. These derived features simplified the data without diluting its meaning. It allowed for focused interpretations such as identifying high-growth cities or flagging anomalies. Visual tools like heatmaps and pie charts enhanced the comprehensibility of complex datasets, enabling faster, intuitive understanding even for non-technical stakeholders.

This project also demonstrates how **open government data can be leveraged** for actionable insights. By making such datasets publicly available, platforms like NDAP enable students, researchers, and policy analysts to contribute to the national knowledge base and offer potential improvements to existing schemes.

In conclusion, the CGHS beneficiary dataset not only tells a story of numbers but reflects the story of evolving healthcare access for central government employees in India. It showcases areas of success, gaps to be addressed, and the dynamic nature of public health systems. With further exploration and integration of more granular data (e.g., age, gender, medical visits), deeper insights can be generated.

This project is a testament to the power of data-driven decision-making in governance. The combination of EDA techniques and clear visual storytelling can empower authorities to make informed choices about where and how to expand the reach and efficiency of healthcare schemes like CGHS.

**Future scope**

While the current analysis of the CGHS beneficiary dataset provides important insights into the distribution and growth trends of central government healthcare users across various cities, there remains significant potential to build upon this foundation. The existing study has opened several avenues for deeper investigation, expansion of analytical frameworks, and incorporation of more dimensions to support strategic decision-making. The following points outline the future scope of this project:

1. Integration with Additional Variables

The current dataset primarily focuses on the number of beneficiaries per city across different years. For a more comprehensive analysis, future datasets can incorporate additional variables such as:

* Demographic data (age, gender, occupation category) of beneficiaries.
* Number of CGHS wellness centres or dispensaries in each city.
* Healthcare utilization patterns, such as frequency of medical visits, common health issues, or types of services availed.
* Budget allocation and expenditure for CGHS across regions.

These additional data points would allow for richer, multi-dimensional analyses that can uncover the impact of infrastructure, demographic trends, and policy changes on healthcare access.

2. Predictive Modelling and Forecasting

Going beyond descriptive analytics, machine learning models can be developed for predictive insights, such as:

* Forecasting future beneficiary numbers using time-series models like ARIMA or Prophet.
* Predicting high-growth or low-growth cities based on historical patterns and external factors.
* Identifying regions at risk of underutilization or overburdening of services.

Such forecasting can aid in proactive resource planning and infrastructure development.

3. City-Level Policy Assessment

Using external datasets such as census data, public health infrastructure indices, and employee distribution data, the analysis can be extended to assess how effectively the CGHS scheme aligns with the healthcare needs of each city. This can inform:

* Policy reforms targeting underserved regions.
* Budget redistribution to areas with growing demand.
* Awareness campaigns in cities with low enrolment.

4. Real-Time Dashboards and Automation

Transforming the static EDA into a real-time or automated dashboard using tools like Tableau, Power BI, or even Python dashboards (e.g., Plotly Dash, Streamlit) can enhance the accessibility and usability of the insights. Key benefits include:

* Live updates as new data becomes available.
* User-interactive filters for dynamic insights.
* Customized views for different stakeholders such as policy-makers, researchers, or citizens.

5. Geographic Visualization and Mapping

Integrating the dataset with geospatial data (e.g., city coordinates, regional maps) would allow the use of GIS (Geographic Information Systems) and heat maps to visually track healthcare coverage and its expansion. This would be particularly effective in:

* Visualizing regional disparities.
* Planning mobile healthcare units.
* Tracking regional effectiveness of CGHS expansions.

6. Comparative Scheme Analysis

The methodology used in this project can be replicated or extended to compare CGHS with other government schemes such as:

* ESIC (Employees' State Insurance Corporation)
* Ayushman Bharat
* State government healthcare programs

A comparative study can reveal overlapping efforts, gaps in service, and opportunities for integration or consolidation.

7. Feedback Loop from Beneficiaries

Incorporating qualitative data such as feedback, satisfaction ratings, or service experience reports from beneficiaries can greatly enhance the understanding of scheme effectiveness. Surveys or sentiment analysis on public forums can also be used to build a more holistic performance model.

8. Expansion to National-Level Planning

Insights gained at the city level can be scaled to aid in national-level healthcare planning. The model can serve as a blueprint for:

* Identifying health service gaps in rural vs urban areas.
* Targeting future CGHS branches.
* Supporting the digital health mission of India.

In conclusion, this project has strong potential to evolve from a foundational EDA into a dynamic decision-support tool for healthcare planning. With additional data, advanced techniques, and collaborative integration with public health bodies, the future scope of this work can contribute meaningfully to policy innovation and efficient healthcare delivery in India.

**References**

[1] NITI Aayog, “CGHS Beneficiaries - Urban,” *National Data and Analytics Platform (NDAP)*, [Online]. Available: <https://ndap.niti.gov.in/dataset/6795>  
[Accessed: April 10, 2025].

[2] Python Software Foundation, “pandas: Python Data Analysis Library,” [Online]. Available: https://pandas.pydata.org/  
[Accessed: April 10, 2025].

[3] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.

[4] M. Waskom et al., “Seaborn: Statistical Data Visualization,” Journal of Open Source Software, vol. 6, no. 60, p. 3021, 2021.

[5] Wes McKinney, Python for DataAnalysis*:* Data Wrangling with Pandas, NumPy, andIPython, 2nd ed., O’Reilly Media, 2017.