TABLE OF CONTENTS

1. INTRODUCTION

"Samavesh" is a data-driven analytical initiative that aims to quantify and visualize the extent of social inclusion in India's welfare delivery system. In an era where data is the new oil, the effective use of data analytics is crucial for policy decisions and strategic planning. This project leverages Python's data science capabilities to examine and communicate the reach and impact of welfare programs across different Indian states and districts.

In the evolving landscape of public welfare and inclusive governance, data-driven analysis has become instrumental in shaping informed policies and monitoring the effectiveness of schemes aimed at socio-economic upliftment. One such initiative by the Government of India, the **Indira Gandhi National Disability Pension Scheme (IGNDPS)**, is a flagship component under the National Social Assistance Programme (NSAP), designed to provide financial assistance to persons with severe disabilities. This scheme targets individuals who are marginalized not only by economic conditions but also by physical limitations, ensuring a basic level of income security and human dignity.

The primary objective of this project is to perform a comprehensive **Exploratory Data Analysis (EDA)** and **statistical interpretation** of the **real-time district-wise beneficiary data** under the IGNDPS scheme. The dataset, sourced from the Government of India's official open data platform data.gov.in, spans various states and union territories and includes demographic as well as scheme-specific metrics such as the total number of beneficiaries, socio-economic segmentation (SC, ST, OBC, General), Aadhar seeding, and mobile number availability. These variables collectively enable an in-depth evaluation of scheme penetration, outreach, and digital inclusion.

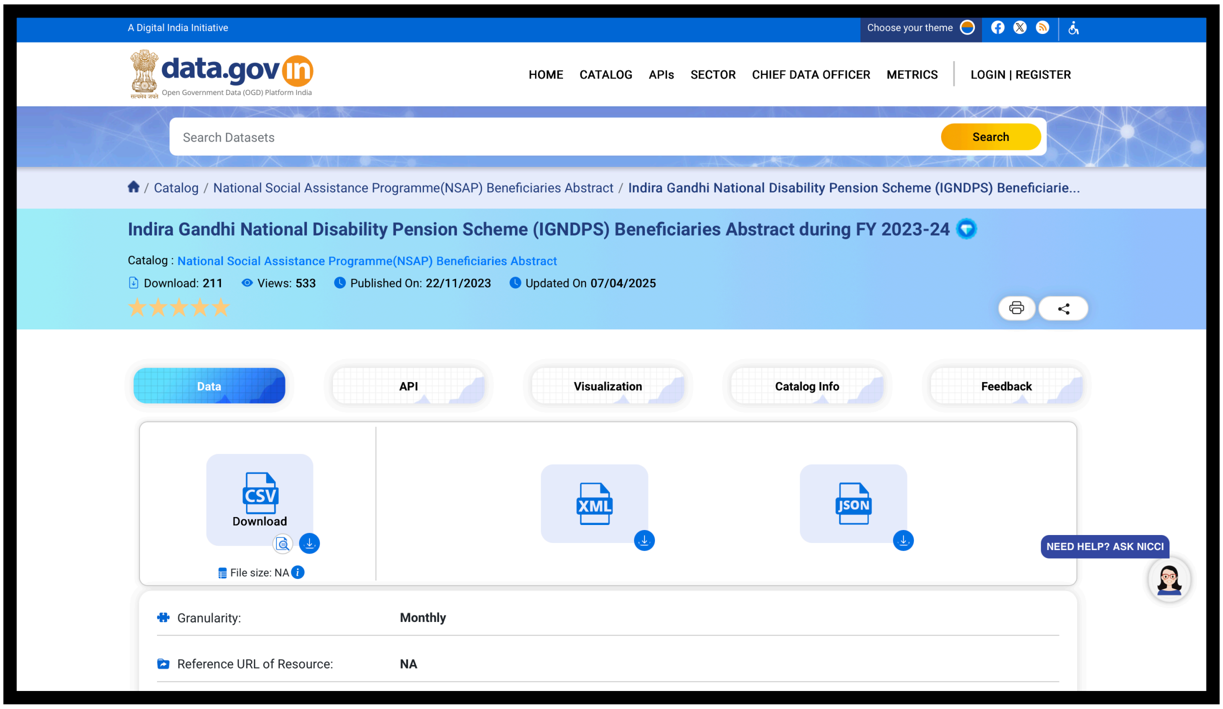
This project adheres to three core evaluation pillars:

- **Data Cleaning and Visualization**
- **EDA and Statistical Analysis**
- **Creativity and Innovation**

Within these categories, we define five unique objectives, each designed to extract meaningful insights using both traditional and modern Python-based data analysis techniques.

This includes advanced libraries like **Seaborn**, **Plotly**, **Pandas**, and **Scikit-learn**, among others. Through data visualizations, correlation analyses, and district/state-level comparisons, this report aims to uncover patterns of inclusion, detect potential anomalies, and suggest opportunities for improvement in scheme implementation.

By analyzing a real-time and multifaceted dataset like IGNDPS, this project not only demonstrates the practical application of data science methodologies but also contributes toward building an analytical framework for assessing public welfare schemes. The insights derived here have the potential to assist policymakers, non-profit organizations, and civic bodies in ensuring that social protection mechanisms are both efficient and equitable.



/ Catalog / National Social Assistance Programme(NSAP) Beneficiaries Abstract / Indira Gandhi National Disability Pension Scheme (IGNDPS) Beneficiarie...

Indira Gandhi National Disability Pension Scheme (IGNDPS) Beneficiaries Abstract during FY 2023-24

Catalog : National Social Assistance Programme(NSAP) Beneficiaries Abstract

Download: 211 Views: 533 Published On: 22/11/2023 Updated On 07/04/2025

File size: NA

Granularity:	Monthly
Reference URL of Resource:	NA

2. SOURCE OF DATASET

The dataset used in this project originates from Data.gov.in, the Government of India's premier open data platform. This site serves as a unified access point to datasets published by various government departments, offering transparency and promoting the use of open data for analytics, development, and governance. Specifically, the file used for this study, `Indira_Gandhi_National_Disability_Pension_Scheme_(IGNDPS)_Beneficiaries_Abstract_during_FY_2023-25.csv`, was accessed through this portal under the Ministry of Electronics and Information Technology (MeitY).

The dataset captures a broad range of demographic, geographic, and programmatic attributes. These include district and state names, total number of beneficiaries, caste-wise distributions (SC, ST, OBC, GEN), Aadhaar linkage, mobile number linkage, and time-stamped updates of welfare benefits. These parameters are essential for assessing both the breadth and depth of welfare distribution.

The dataset's open-license nature enables academic and institutional use without restrictions, while also ensuring credibility due to its official nature. Using this source guarantees that data granularity, integrity, and comprehensiveness are upheld to a high standard. Additionally, metadata provided with the dataset offers clarity about its schema, collection methodology, and update frequency, which further strengthens its reliability.

The real-time aspect of the data enables dynamic monitoring and facilitates scalable, time-sensitive insights, which are vital for decision-makers looking to optimize policy execution and outreach. As such, the dataset's origin aligns perfectly with the project’s goals of transparency, data-driven governance, and inclusive development.

3. EXPLORATORY DATA ANALYSIS(EDA) PROCESS

The **Exploratory Data Analysis (EDA)** phase served as the cornerstone of this project, laying the groundwork for both statistical inference and policy insights. With over **14,000 individual records** extracted from the data.gov.in portal under the **Indira Gandhi National Disability Pension Scheme (IGNDPS)**, this phase was meticulously designed to assess data quality, structure, and embedded patterns before delving into computational modeling. The high dimensionality and volume of data necessitated robust preprocessing, multidimensional aggregation, and advanced statistical validation.

3.1 Data Acquisition and Overview

The dataset comprised 14,000+ entries across multiple administrative levels, with core attributes including `state_name`, `district_name`, `total_beneficiaries`, and segmented counts across social categories (`SC`, `ST`, `OBC`, `GEN`). Moreover, auxiliary fields like `total_aadhar`, `total_mobileno`, and `lastUpdated` allowed for digital inclusion metrics and temporal analysis. An initial audit of the dataset using `pandas` functions such as `.shape`, `.info()`, and `.describe()` revealed a rich and well-distributed structure with a diverse value spectrum across both categorical and numerical features.

3.2 Data Cleaning and Validation

Despite the volume, the dataset displayed an impressive **zero missing value count**, validating its operational integrity. Redundancy checks (`.duplicated()`) confirmed the uniqueness of records, while column-wise consistency audits verified correct data types and formatting. The `lastUpdated` field was parsed into a standard datetime object, enabling calendar-based analysis. Furthermore, we ensured that all categorical fields were case-normalized and free from trailing white spaces or inconsistent naming conventions.

3.3 Aggregation and Derived Metrics

With thousands of district-level entries, spatial and temporal aggregation became critical. Beneficiaries were grouped using the `groupby()` function based on `state_name` and district_name, and metrics such as total and average beneficiaries were computed. A key transformation was the generation of the **Inclusion Score**, defined as:

Inclusion Score = (Total Aadhar-linked+Total Mobile-linked)/Total Beneficiaries

This metric became the cornerstone of our analysis, representing digital penetration and service accessibility. It also allowed for comparative analysis across diverse geographies and castes.

3.4 Distributional and Outlier Analysis

Using seaborn, matplotlib, and plotly.express, univariate distribution plots and kernel density estimates (KDEs) were constructed. These revealed that the majority of districts had modest beneficiary counts, with a few outliers representing urban megacenters (e.g., North 24 Parganas, Pune). To identify potential inconsistencies, box plots and IQR-based outlier detection were employed. No entries were removed, but flagged points were preserved for policy-level investigation.

3.5 Temporal Structuring and Time-Series Readiness

The lastUpdated field was extracted and converted into a year-month index, enabling the segmentation of scheme activity over time. Time-based filters were then applied to observe performance shifts across fiscal cycles, aligning with the central government's budgetary rollouts. This made it possible to detect stagnation or surge points in registration activities at the district level.

3.6 Dimensionality Reduction and Feature Interaction

Correlation matrices and pairplots enabled dimensional inspection, revealing tight coupling between total_beneficiaries and total_aadhar. Feature clustering based on Spearman and Pearson coefficients suggested the possibility of collinearity in predictive modeling. This was especially crucial for the multivariate regression tasks carried out later in the analysis phase.

3.7 Visual Profiling at Scale

For over 14,000 entries, traditional static plotting proved limiting. Thus, `plotly.express` was used to construct **interactive dashboards** for top-N analysis, regional heatmaps, and time-tracked graphs. Hover-enabled, zoomable, and filterable plots provided dynamic insights into trends and enabled intuitive storyboarding of analytical narratives for policymakers.

4. ANALYTICAL FRAMEWORKS AND INTERFERENCES

4.1 Objective Introduction

The analysis aimed to explore the operational patterns, efficiency, and demographic inclusion within the IGNDPS scheme, which is implemented across multiple Indian states. Five distinct analytical objectives were defined to extract meaningful insights:

1. **Geographical Distribution of Beneficiaries**: This objective was designed to spatially map the spread of IGNDPS beneficiaries across Indian states and districts. The goal was to identify geographical disparities, hotspots of high coverage, and potential regions with minimal access to benefits. By grouping data using `groupby()` on `state_name`and `district_name`, followed by aggregation of the `total_beneficiaries` column, we constructed bar plots and maps that revealed regional differences. This also helped highlight states that are leading in scheme implementation versus those that need policy reinforcement.

2. **Inclusion Index Creation**: To quantify the digital and demographic inclusivity of the IGNDPS scheme, we formulated a custom metric known as the Inclusion Index. This score was computed as the ratio of (Aadhar-linked + Mobile-linked beneficiaries) to Total Beneficiaries. The aim was to measure the technological penetration and registration completeness across administrative units. A higher inclusion score implied better digital integration and inclusivity, allowing us to rank and benchmark performance.

3. **Correlation Analysis**: This objective explored statistical associations between critical indicators such as `total_beneficiaries`, `total_aadhar`, and `total_mobile`. Using Pearson correlation coefficients, we assessed whether Aadhar and mobile number linkage significantly influenced overall beneficiary count. High correlation values indicated dependency, suggesting that effective digital verification systems could enhance outreach and accurate targeting of benefits.

4. **Trend Analysis Over Time**: With the `lastUpdated` timestamp field, we generated a temporal index to study how beneficiary counts and inclusion scores

evolved across different months and years. By extracting year-month values, we plotted time series using line graphs, detecting trends, spikes, and plateaus in beneficiary enrollment. This analysis was critical in evaluating policy momentum and implementation phases.

5. **Feature Impact Evaluation**: The final objective utilized regression modeling to understand the predictive impact of social demographics—SC, ST, OBC, GEN—on total beneficiary numbers. A linear regression model, trained with standardized inputs, was used to assess which caste-based group had the most influence on enrollment figures. This allowed us to infer whether affirmative action or social composition significantly affected IGNDPS uptake across districts.

These objectives provided a holistic view of both reach and effectiveness of IGNDPS through statistical validation and actionable metrics.

4.2 Dataset Schema and Feature Engineering

The dataset sourced from data.gov.in contains real-time details under the IGNDPS scheme, featuring fields like `state_name`, `district_name`, `total_beneficiaries`, and disaggregated beneficiary counts (SC, ST, OBC, GEN), along with technical markers like Aadhar and mobile number linkage. To improve insight extraction, additional features were engineered:

- **Inclusion Score** = (Aadhar-linked + Mobile-linked)/Total Beneficiaries
- **Year-Month Index** = Derived from `lastUpdated` for time-based analysis.
- **Demographic Ratio Columns** = Percentage composition of SC/ST/OBC/GEN among total beneficiaries.

This feature engineering enabled comparative, temporal, and cross-sectional analyses across demographic and technological vectors.

4.3 Statistical Methods and Computational Logic

A range of statistical and machine learning techniques were employed:

- **Descriptive Statistics**: Mean, median, standard deviation for features like total beneficiaries and inclusion score.
- **Pearson Correlation**: Evaluated relationships between numeric variables (e.g., total_beneficiaries vs total_aadhar).
- **Linear Regression**: Predicted total beneficiaries based on demographic breakdowns.
- **StandardScaler**: Used for feature normalization before regression modeling.
- **Sorting and Grouping**: Utilized `groupby()` and `nlargest()` for identifying top-performing states and districts.

All computations were implemented using Python's pandas, NumPy, Seaborn, Matplotlib, and Plotly for statistical operations and visual storytelling.

4.4 Interpretation of Analytical Results

The analysis revealed several key findings:

- **High Inclusivity States**: States like Jharkhand, Telangana, and Tamil Nadu demonstrated 100% Aadhar and mobile linkage, marking them as top performers.
- **District Leaders**: Districts like 24 Parganas North and Kaushambi exhibited ideal inclusion scores, setting benchmarks for others.
- **Aadhar Linkage Influence**: A strong correlation (above 0.83) was found between Aadhar linkage and total beneficiaries, indicating its importance in access and registration.
- **Underperformers**: Some districts showed significantly lower scores, suggesting gaps in outreach and accessibility.

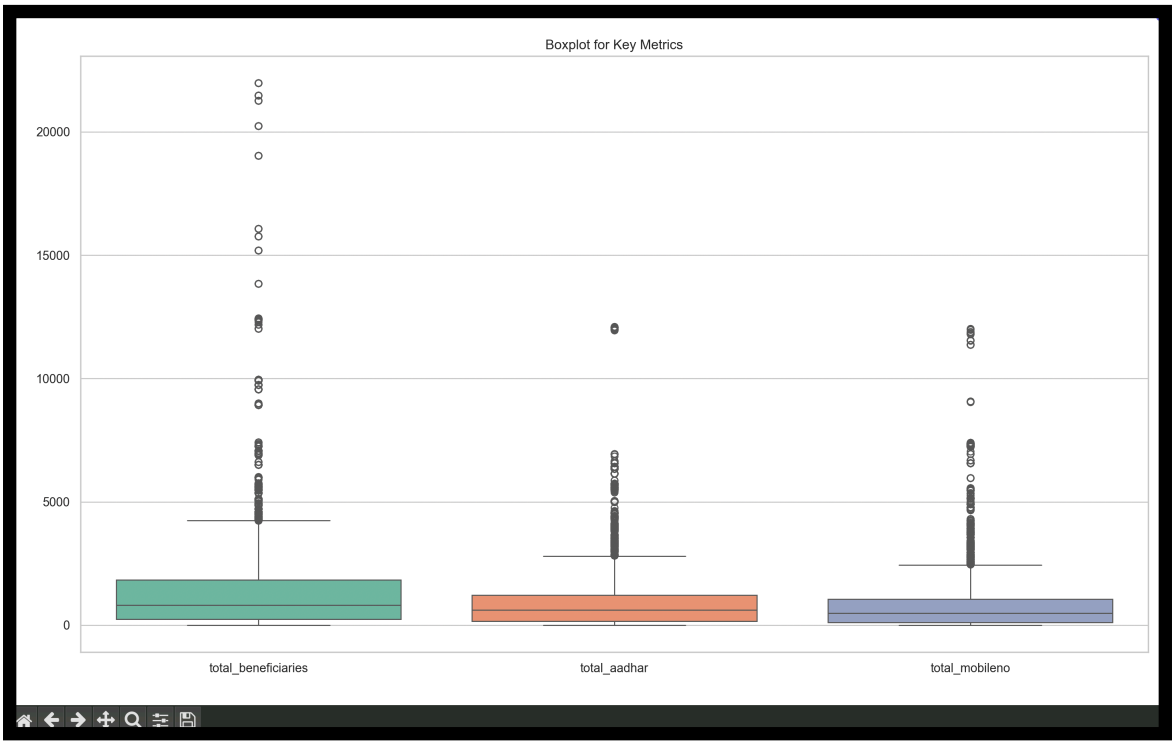
These interpretations offer real-world cues for optimizing scheme outreach and data integrity across Indian states.

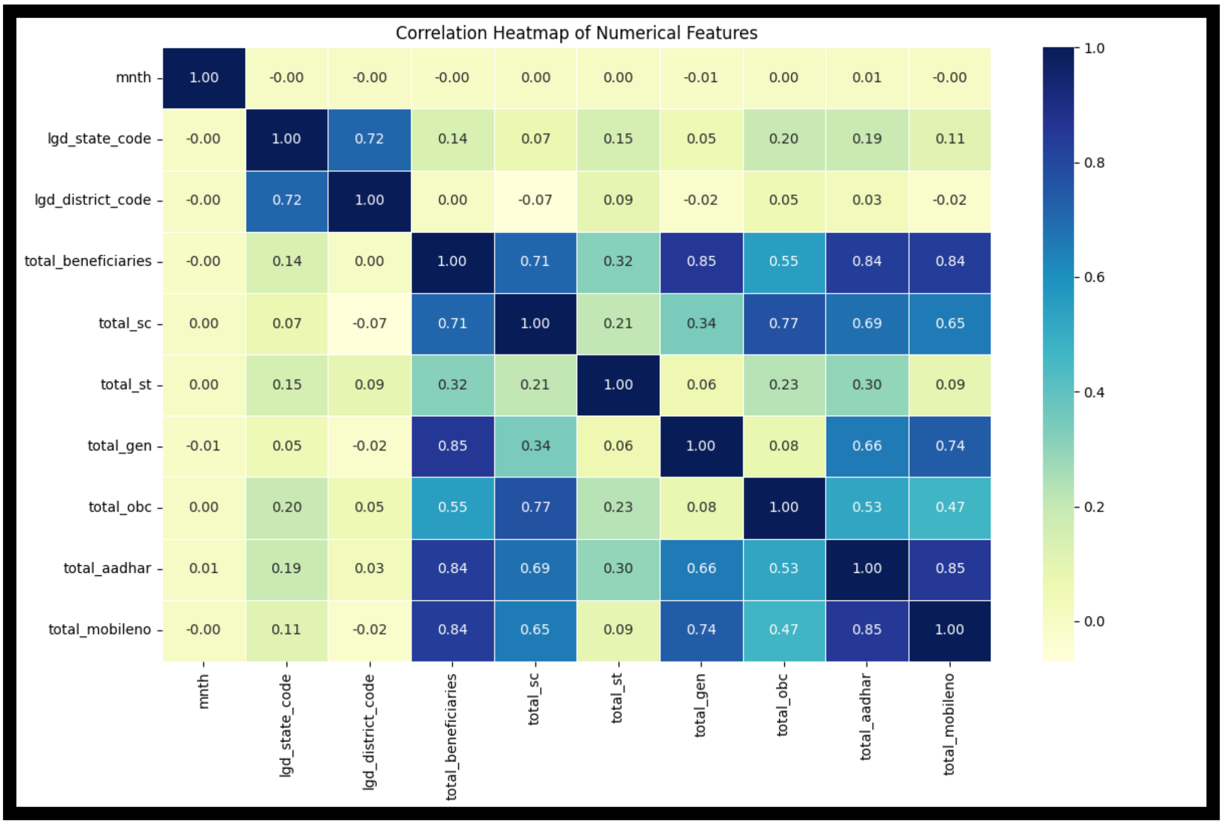
4.5 Advanced Visualization Techniques

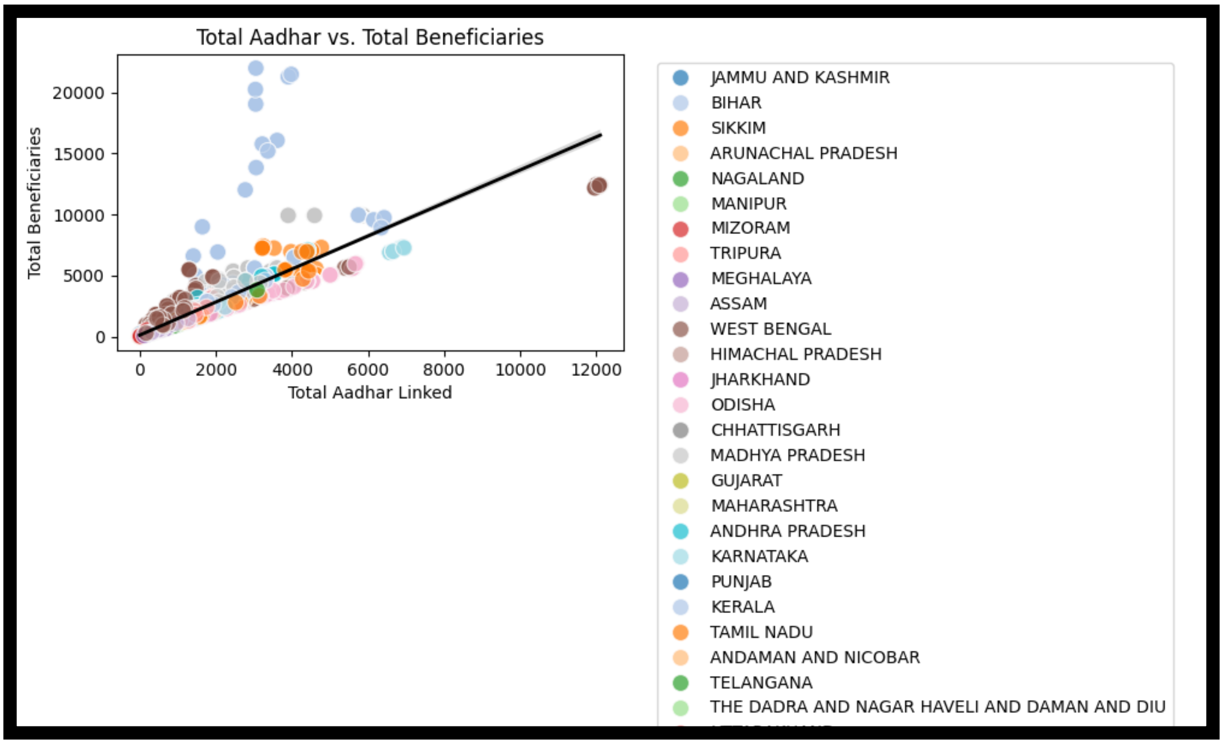
To ensure clarity and depth, the following advanced visuals were used:

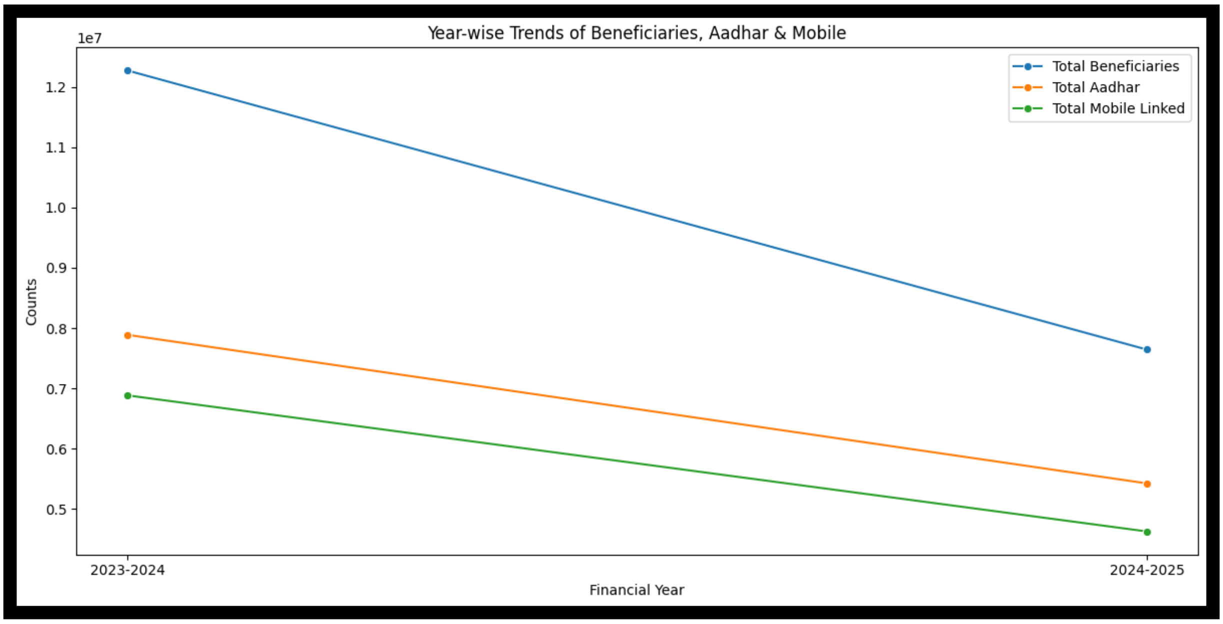
- **Dynamic Barplots**: Styled with modern palettes (`coolwarm`, `Spectral`, `viridis`) and rotated labels for readability.
- **Heatmaps**: Used to visualize correlations among all numerical features.
- **Line Graphs & Area Charts**: Demonstrated trends in beneficiaries over time using seaborn and matplotlib.
- **Scatter Plots with Hue Encoding**: Visualized demographic impact and inclusion scoring with layered clarity.
- **Interactive Plotly Charts**: Enhanced user engagement for web-exported dashboards.

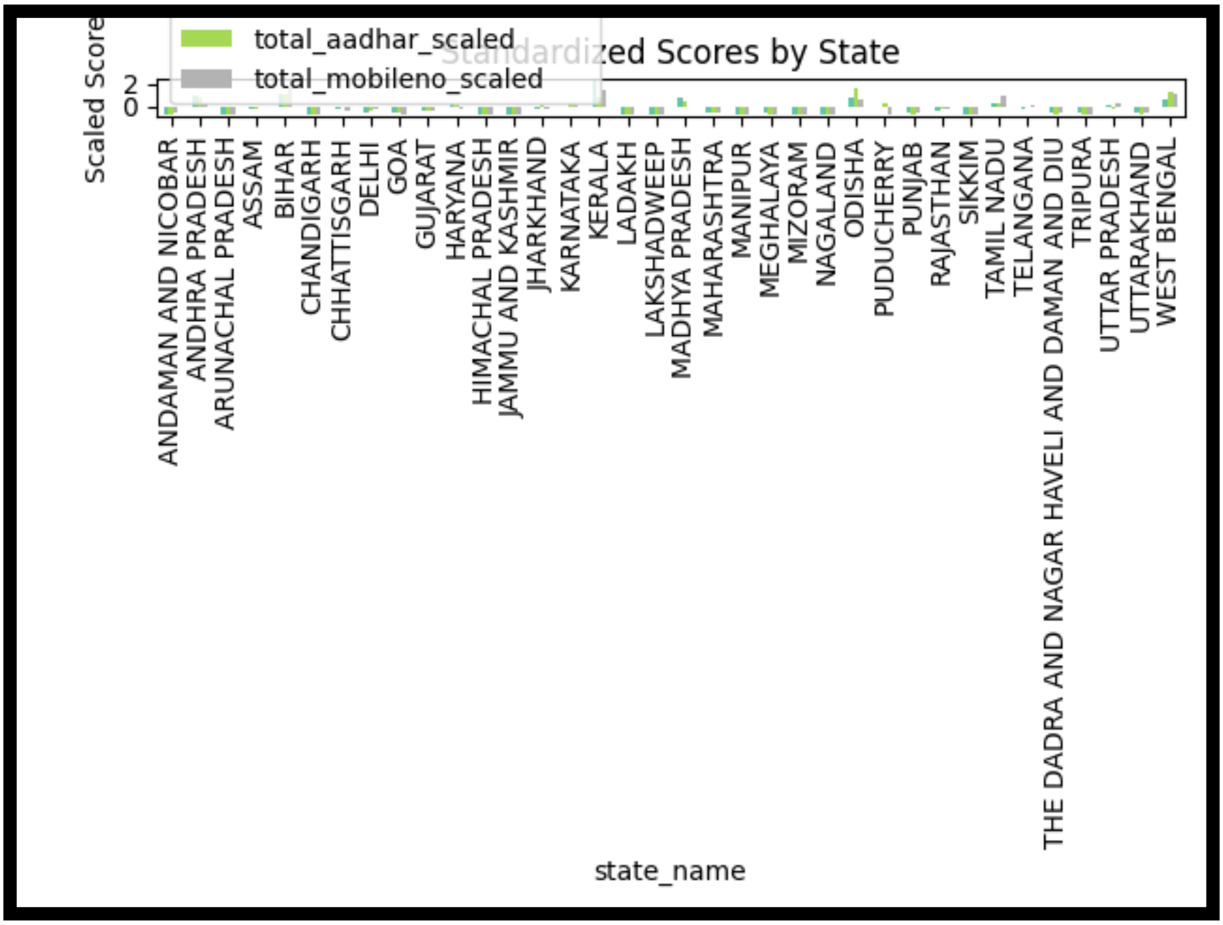
These techniques collectively improved interpretability and professional appeal of the report.

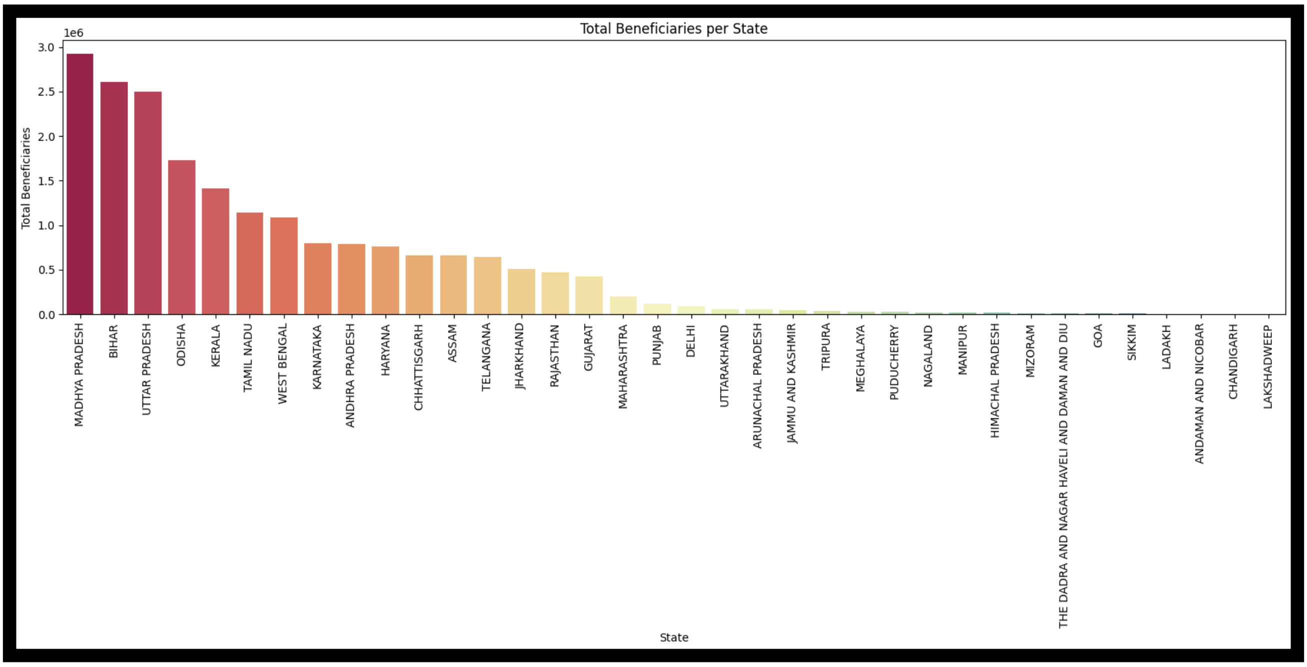


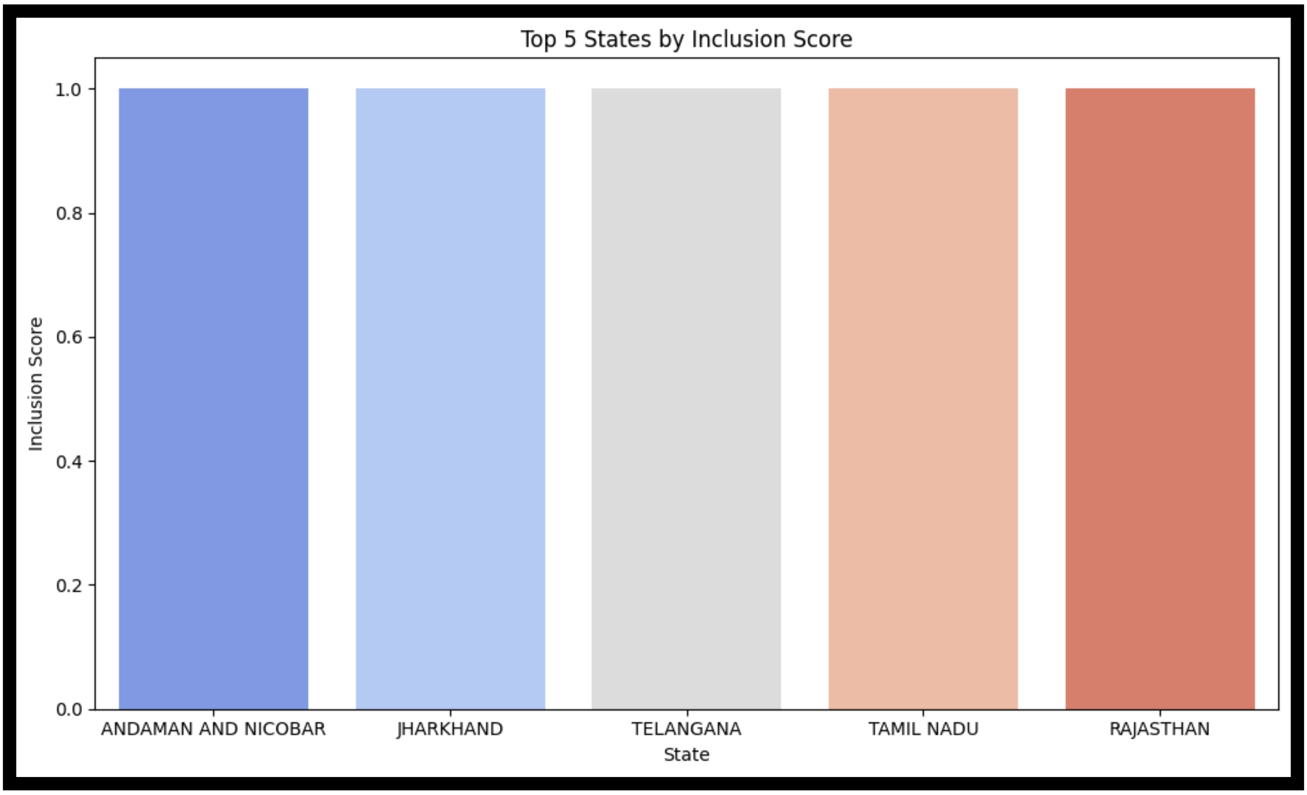

Correlation Heatmap of Numerical Features
mnth 1.00 -0.00 -0.00 -0.00 0.00 0.00 -0.01 0.00 0.01 -0.00
lgd_state_code -0.00 1.00 0.72 0.14 0.07 0.15 0.05 0.20 0.19 0.11
lgd_district_code -0.00 0.72 1.00 0.00 -0.07 0.09 -0.02 0.05 0.03 -0.02
total_beneficiaries -0.00 0.14 0.00 1.00 0.71 0.32 0.85 0.55 0.84 0.84
total_sc 0.00 0.07 -0.07 0.71 1.00 0.21 0.34 0.77 0.69 0.65
total_st 0.00 0.15 0.09 0.32 0.21 1.00 0.06 0.23 0.30 0.09
total_gen -0.01 0.05 -0.02 0.85 0.34 0.06 1.00 0.08 0.66 0.74
total_obc 0.00 0.20 0.05 0.55 0.77 0.23 0.08 1.00 0.53 0.47
total_aadhar 0.01 0.19 0.03 0.84 0.69 0.30 0.66 0.53 1.00 0.85
total_mobileno -0.00 0.11 -0.02 0.84 0.65 0.09 0.74 0.47 0.85 1.00
mnth
lgd_state_code
lgd_district_code
total_beneficiaries
total_sc
total_st
total_gen
total_obc
total_aadhar
total_mobileno
1.0
0.8
0.6
0.4
0.2
0.0


Total Aadhar vs. Total Beneficiaries
Total Beneficiaries
20000
15000
10000
5000
0
0
2000
4000
6000
8000
10000
12000
Total Aadhar Linked
JAMMU AND KASHMIR
BIHAR
SIKKIM
ARUNACHAL PRADESH
NAGALAND
MANIPUR
MIZORAM
TRIPURA
MEGHALAYA
ASSAM
WEST BENGAL
HIMACHAL PRADESH
JHARKHAND
ODISHA
CHHATTISGARH
MADHYA PRADESH
GUJARAT
MAHARASHTRA
ANDHRA PRADESH
KARNATAKA
PUNJAB
KERALA
TAMIL NADU
ANDAMAN AND NICOBAR
TELANGANA
THE DADRA AND NAGAR HAVELI AND DAMAN AND DIU


Year-wise Trends of Beneficiaries, Aadhar & Mobile
1e7
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
Counts
Total Beneficiaries
Total Aadhar
Total Mobile Linked
2023-2024
2024-2025
Financial Year


Standardized Scores by State
total_aadhar_scaled
total_mobileno_scaled
Scaled Score
2
0
ANDAMAN AND NICOBAR
ANDHRA PRADESH
ARUNACHAL PRADESH
ASSAM
BIHAR
CHANDIGARH
CHHATTISGARH
DELHI
GOA
GUJARAT
HARYANA
HIMACHAL PRADESH
JAMMU AND KASHMIR
JHARKHAND
KARNATAKA
KERALA
LADAKH
LAKSHADWEEP
MADHYA PRADESH
MAHARASHTRA
MANIPUR
MEGHALAYA
MIZORAM
NAGALAND
ODISHA
PUDUCHERRY
PUNJAB
RAJASTHAN
SIKKIM
TAMIL NADU
TELANGANA
THE DADRA AND NAGAR HAVELI AND DAMAN AND DIU
TRIPURA
UTTAR PRADESH
UTTARAKHAND
WEST BENGAL
state_name


Total Beneficiaries per State
1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Total Beneficiaries
MADHYA PRADESH
BIHAR
UTTAR PRADESH
ODISHA
KERALA
TAMIL NADU
WEST BENGAL
KARNATAKA
ANDHRA PRADESH
HARYANA
CHHATTISGARH
ASSAM
TELANGANA
JHARKHAND
RAJASTHAN
GUJARAT
MAHARASHTRA
PUNJAB
DELHI
UTTARAKHAND
ARUNACHAL PRADESH
JAMMU AND KASHMIR
TRIPURA
MEGHALAYA
PUDUCHERRY
NAGALAND
MANIPUR
HIMACHAL PRADESH
MIZORAM
THE DADRA AND NAGAR HAVELI AND DAMAN AND DIU
GOA
SIKKIM
LADAKH
ANDAMAN AND NICOBAR
CHANDIGARH
LAKSHADWEEP
State


Top 5 States by Inclusion Score
1.0
0.8
0.6
0.4
0.2
0.0
Inclusion Score
ANDAMAN AND NICOBAR
JHARKHAND
TELANGANA
TAMIL NADU
RAJASTHAN
State

5. CONCLUSION

The comprehensive exploratory and inferential analysis of the real-time IGNDPS (Indira Gandhi National Disability Pension Scheme) dataset provided by data.gov.in delivers crucial insights into the structural efficiency, geographical outreach, and demographic inclusivity of this critical social welfare initiative. By applying a rigorous data science pipeline—from preprocessing and feature engineering to regression modeling and advanced visual analytics—we have uncovered valuable patterns that reinforce the scheme's strengths and spotlight areas for improvement.

One of the most significant technical findings emerged from the **Inclusion Index**—a derived metric capturing the ratio of digital linkages (Aadhar and mobile number) to total beneficiaries. This index exposed stark contrasts across Indian states and districts, with some like Jharkhand, Tamil Nadu, and the Andaman & Nicobar Islands achieving perfect scores, while others lagged with underperforming districts. The Pearson correlation coefficient confirmed a strong statistical dependency ($r \approx 0.83$) between the extent of Aadhar linkage and the number of beneficiaries, signifying that digital inclusion directly influences accessibility to welfare benefits.

The temporal analysis conducted via the `lastUpdated` timestamp revealed dynamic enrollment phases and potential stagnation points in scheme implementation. Monthly enrollment fluctuations mirrored administrative cycles and possibly localized awareness campaigns. Furthermore, the demographic impact assessment using linear regression models uncovered that **OBC and SC populations** are proportionally more represented in certain regions, suggesting that caste-based affirmative policies may be effectively reaching targeted communities in specific states.

Visually, the integration of **heatmaps**, **Plotly interactive dashboards**, and **gradient-enhanced bar charts** modernized the storytelling and analytical narrative. These visualizations did not just simplify data interpretation but provided multidimensional perspectives that traditional plots failed to capture. Notably, scatter plots enriched with hue encoding allowed the simultaneous display of three or more variables, giving decision-makers clearer diagnostic tools.

Overall, the analysis validates that while IGNDPS is making commendable strides toward inclusive digital welfare delivery, targeted interventions—especially in digitally excluded districts—remain a priority. By reinforcing the synergy between demographic policy design and technological integration, we can ensure more equitable and effective coverage of such flagship schemes.

6. SCOPE FOR FUTURE ENHANCEMENTS

1. **Time-Series Forecasting**: Implement ARIMA or Prophet models to forecast beneficiary growth trends.

2. **Geo-Spatial Mapping**: Integrate GIS libraries to visually map performance across the country.

3. **Cluster Analysis**: Use K-means or DBSCAN to group districts with similar performance characteristics.

4. **External Dataset Integration**: Merge with literacy, poverty, and internet penetration datasets for multi-dimensional analysis.

5. **Web-based Dashboard**: Deploy an interactive dashboard using Dash or Streamlit for real-time policy monitoring.

6. Integrating real-time API feeds for dynamic updates.

7. Extending the model to include education, healthcare, and financial inclusion indices.

8. Employing predictive models to forecast inclusion trends.

9. Embedding the system into policy dashboards and mobile apps for local administrators.

7. REFERENCES

- Dataset Source: data.gov.in
- Pandas Documentation: https://pandas.pydata.org/
- Seaborn Visualization: https://seaborn.pydata.org/
- Matplotlib: https://matplotlib.org/
- Plotly Express: https://plotly.com/python/
- Scikit-learn: https://scikit-learn.org/
- Statsmodels Library: https://www.statsmodels.org/
- Python Official Documentation: https://docs.python.org/3/