UIDAI Data Hackathon 2026

Hackathon Submission Report

Problem Theme

Unlocking Societal Trends in Aadhaar Enrolment and Updates

Project Title

UDHAI – National MSME Analytics & Decision Support Portal

Institution / Department

Department of Artificial Intelligence

Vishwakarma University, Pune, Maharashtra, 411048

By Team ArthNiti Analytics

Parvesh Jain

Vandith Shetty

Gagan Agrawal

GitHub Link -- https://github.com/praveshjainnn/UIDAI-Data-Hackathon-2026

Submitted By ArthNiti Analytics Team

1. Problem Statement and Approach

1.1 Background and Context

Aadhaar-linked enrolment and update data forms the backbone of several governance and welfare systems in India. In the context of Micro, Small, and Medium Enterprises (MSMEs), Aadhaar integration through the Udyam registration process has created a large, structured administrative dataset that connects enterprise activity with demographic, geographic, and socio-economic attributes.

MSMEs play a critical role in India's economy by contributing to employment generation, regional development, and social inclusion. However, despite the availability of Aadhaar-linked MSME data, its usage has largely remained limited to registration, verification, and compliance purposes. The deeper societal trends embedded within this data, such as regional disparities, inclusion gaps, employment efficiency, and sectoral balance, remain underexplored.

Unlocking these trends is essential for moving from generic policy formulation to evidence-based, targeted interventions. A structured analytical approach can convert Aadhaar enrolment data from a static administrative record into a dynamic tool for understanding how entrepreneurship, employment, and inclusion evolve across regions and communities.

1.2 Problem Statement

Although Aadhaar-linked MSME enrolment data is extensive and regularly updated, policymakers and administrators face several limitations in its current form:

- Regional MSME growth patterns are difficult to compare due to lack of integrated geographic analysis

- Social inclusion indicators such as gender and caste participation are not systematically analysed
- Employment outcomes are not evaluated in relation to investment or enterprise scale
- There is no unified framework to benchmark states or districts based on MSME ecosystem development

As a result, decision-making often relies on aggregated summaries that fail to capture multi-dimensional societal trends.

The central problem addressed in this project is:

How can Aadhaar-linked MSME enrolment data be systematically analysed and visualised to uncover societal, regional, and economic trends that support informed policy and administrative decision-making?

1.3 Objectives of the Study

The objectives of this project are:

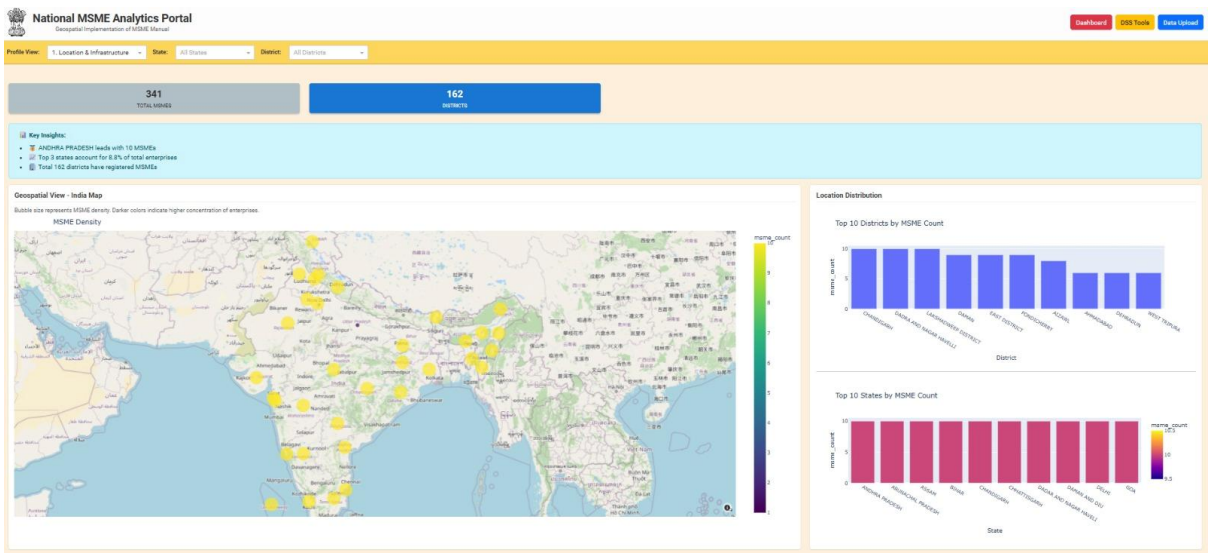
- To analyse Aadhaar-linked MSME data across geographic, social, economic, and sectoral dimensions
- To identify regional and societal disparities in MSME participation and outcomes
- To evaluate employment generation efficiency across states and sectors
- To develop a comparative framework for benchmarking MSME ecosystem development
- To present insights through clear, interpretable visualisations suitable for administrative use

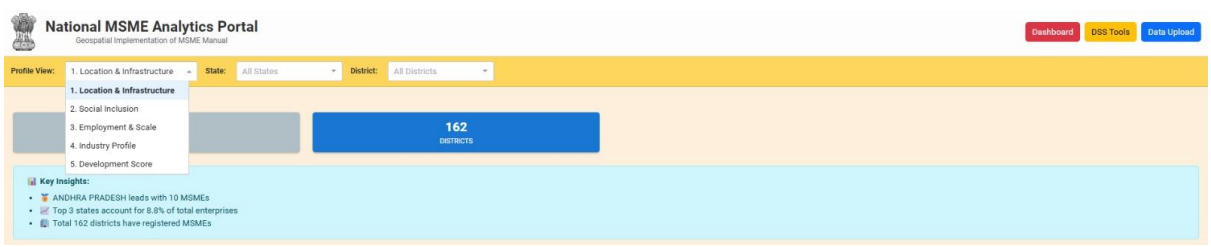
1.4 Proposed Analytical Approach

The project adopts a structured, data-driven approach consisting of four integrated components:

1. **Multi-Dimensional Profiling**
   MSME data is analysed across location, social inclusion, employment, investment, and industry dimensions to capture diverse societal trends.
2. **Exploratory and Statistical Analysis**
   Univariate, bivariate, and trivariate analyses are applied to understand distributions, relationships, and interactions between key variables.
3. **Composite Development Scoring**
   A transparent composite index is designed to benchmark states based on MSME scale, social inclusion, employment generation, and industry diversity.
4. **Interactive Visualisation and Decision Support**
   Analytical findings are presented through dashboards and geospatial visualisations that allow users to explore patterns, identify outliers, and support data-driven decisions.

This approach ensures that Aadhaar-linked MSME enrolment data is transformed into **actionable insights** rather than static statistics.





Submitted By ArthNiti Analytics Team

2. Datasets Used

2.1 Primary Dataset Description

The analysis is based on the Aadhaar-linked MSME (Udyam) registration dataset provided by UIDAI for the hackathon. This dataset contains anonymized enterprise-level records where Aadhaar enrolment is used as the primary identity mechanism during MSME registration and updates.

The dataset enables the study of societal trends by linking demographic attributes with enterprise characteristics such as employment, investment, and sectoral classification. It provides coverage across multiple states and districts in India, making it suitable for geographic and comparative analysis.

The dataset represents a snapshot of registered MSMEs during the Udyam registration period and is treated as an administrative dataset intended for analytical and policy-oriented use rather than individual-level tracking.

2.2 Dataset Attributes and Usage

The following categories of variables were used in the analysis:

Geographic Attributes

- State and district identifiers were used as the primary aggregation keys for regional analysis
- PIN code information supported finer-grained locality-level grouping where required

Demographic and Social Attributes

- Gender information enabled analysis of female participation in entrepreneurship
- Social category indicators (such as SC, ST, OBC, and General) were used to assess inclusivity

- Disability status was used to study participation of differently-abled entrepreneurs

Enterprise Characteristics

- Enterprise type (Micro, Small, Medium) was used to study scale distribution
- Organization type supported structural analysis of ownership patterns

Economic Indicators

- Total employment was used to measure job creation outcomes
- Investment amount was used to assess capital intensity and efficiency

Sectoral Classification

- NIC codes were used to categorize enterprises into manufacturing and services sectors
- Industry classification enabled analysis of sectoral diversity and balance

Each variable was selected based on its relevance to understanding societal, economic, or regional trends in Aadhaar-linked MSME enrolment.

Aid	Enterprisename	Socialcategory	Gender	Ph		Organisationtyp	Plantlocation	Address	State	District	Pincode	Commmenced
8017346	ACCOMBLISS PR	General	Male	No	Private Limited C	1) SHOP NO 14	SHOP NO 146, S	CHANDIGARH	CHANDIGARH	160036	20-02-2017	Services
8025200	Smile Every Mile	General	Male	No	Proprietary	1) SCO 2441-24	Sco 2441-2442, I	CHANDIGARH	CHANDIGARH	160022	03-08-2018	Services
8029011	HOME WORK FA	General	Male	No	Proprietary	1) HOUSE NO. -	HOUSE NO. 392	CHANDIGARH	CHANDIGARH	160025	01-07-2017	Services
8046790	H R ENTERPRISE	General	Male	No	Proprietary	1) SCO 37 SCO	SCO 37 SECTOR	CHANDIGARH	CHANDIGARH	160036	15-01-2016	Manufacturing
8045392	COGNITIVE HEA	General	Male	No	Partnership	1) SHOP NO 3 F	SHOP NO 3, PLC	CHANDIGARH	CHANDIGARH	160014	03-05-2017	Manufacturing
7981150	HOTEL R.R VILLA	General	Male	No	Proprietary	1) PLOT NO. 51	HOTEL R.R.VILLA	CHANDIGARH	CHANDIGARH	160102	08-02-2019	Services
8007103	DHAN LUXMI N	General	Male	No	Proprietary	1) 1020/ B DHA	1020/ B,VILLAGE	CHANDIGARH	CHANDIGARH	160101	04-02-2014	Manufacturing
8041867	M/s FITDRILL EN	General	Male	No	Partnership	1) HOUSE NO. E	HOUSE NO. 304	CHANDIGARH	CHANDIGARH	160019	11-11-2019	Services
8042783	APOSTROPHE A	General	Male	No	Public Limited C	1) House numb	House Number-	CHANDIGARH	CHANDIGARH	160002	17-01-2008	Services
8045193	SIPRO ENTERPRI	OBC	Female	No	Proprietary	1) SCO.189/1 M	SCO.189/1 , MO	CHANDIGARH	CHANDIGARH	160047	01-04-2018	Services
8086781	Nasreen Enterpr	General	Male	No	Proprietary	1) Hubby colon	Hubby colony, B	JAMMU AND KA	SRINAGAR	190023	15-10-2019	Manufacturing
8088630	Knowracle Publi	General	Female	No	Proprietary	1) Shop No. 1 S	House No. 15 La	JAMMU AND KA	JAMMU	180012	01-02-2016	Services
7980599	S.KJINDUSTRIES	General	Male	No	Proprietary	1) - IGC, PHASE	IGC, PHASE III, S	JAMMU AND KA	SAMBA	184121	09-11-2019	Manufacturing
8006832	MAKHAN CLOTH	General	Male	No	Proprietary	1) 8 VILL-MASH	vill-Mashka,po-H	JAMMU AND KA	KATHUA	184201	01-04-2019	Services
8006002	M/S PERFECT PC	General	Male	No	Partnership	1) PLOT NO. 85	PLOT NO 85, PH	JAMMU AND KA	JAMMU	180010	01-06-2007	Manufacturing
8045044	AARAV TRADERS	General	Male	No	Proprietary	1) DIANI SAMBA	DIANI, SAMBA, S	JAMMU AND KA	SAMBA	184121	18-11-2019	Services
8084080	M/S MUSKAN TI	General	Female	No	Proprietary	1) KH NO 445 M	KH NO 445, MO	JAMMU AND KA	JAMMU	180010	24-09-2015	Manufacturing
8037256	GUPTA INDUSTR	General	Male	No	Partnership	1) 116 PHASE -	116, PHASE - II,	JAMMU AND KA	JAMMU	180010	01-04-2000	Manufacturing
8039260	NUSU ENTERPRI	ST	Male	No	Partnership	1) 216 TRESPON	C/O BSNL OFFIC	JAMMU AND KA	SRINAGAR	194103	01-11-2019	Services
8057733	HEALTHWAY DIA	General	Male	No	Proprietary	1) 0 HEALTHWA	HEALTHWAY DIA	JAMMU AND KA	SRINAGAR	190024	01-11-2019	Services
265882	Nagar vehicle	OBC	NA	NA	Proprietary	NA	Gram Barkheda	MADHYA PRADE	RAJGARH	465689	21/04/2015	Services
265893	SURESH TRAVEL	General	NA	NA	Proprietary	NA	BEGAMPURA UJ	MADHYA PRADE	UJJAIN	456001	12/01/2016	Services
265904	BHAGIRATH FLO	OBC	NA	NA	Proprietary	NA	HOUSE NO.409	MADHYA PRADE	DEWAS	455001	07/08/2015	Manufacturing
265920	dairy products	OBC	NA	NA	Proprietary	NA	Gram Post Natar	MADHYA PRADE	RAJGARH	465689	23/04/2015	Manufacturing
265964	MARMAT TRAVE	General	NA	NA	Proprietary	NA	NARAYAN PURA	MADHYA PRADE	UJJAIN	456001	08/01/2016	Services

Submitted By ArthNiti Analytics Team

Majoractivity	Enterprisetype	Nic5digitcode	Totalemp	Investmentcost	Dic_name	Registrationdat	Lg_dist_code
Small	1) 90001; 2) 900	5	25	CHANDIGARH	15-11-2019	44	
Micro	1) 79110	2	1	CHANDIGARH	16-11-2019	44	
Micro	1) 82990; 2) 749	5	1	CHANDIGARH	16-11-2019	44	
Micro	1) 13999	6	10	CHANDIGARH	19-11-2019	44	
Micro	1) 32909	8	25	CHANDIGARH	19-11-2019	44	
Micro	1) 55101	2	4	CHANDIGARH	11-11-2019	44	
Micro	1) 10712; 2) 107	4	4	CHANDIGARH	14-11-2019	44	
Micro	1) 85410; 2) 960	2	1	CHANDIGARH	18-11-2019	44	
Micro	1) 68100	7	10	CHANDIGARH	18-11-2019	44	
Micro	1) 74909	1	1	CHANDIGARH	19-11-2019	44	
Micro	1) 10795; 2) 107	6	2	SRINAGAR	23-11-2019	13	
Small	1) 58111	8	25	JAMMU	23-11-2019	5	
Micro	1) 24109; 2) 310	8	25	SAMBA	11-11-2019	624	
Micro	1) 96091	1	5	KATHUA	14-11-2019	7	
Small	1) 16109	15	97	JAMMU	14-11-2019	5	
Micro	1) 85500; 2) 821	1	3	SAMBA	19-11-2019	624	
Micro	1) 13924; 2) 139	4	15	JAMMU	23-11-2019	5	
Micro	1) 13924; 2) 139	4	18	JAMMU	18-11-2019	5	
Micro	1) 95210	2	3	SRINAGAR	18-11-2019	13	
Micro	1) 86100	2	10	SRINAGAR	20-11-2019	13	
Micro	1) 47739	2	1	RAJGHAR	24/02/2016	422	
Micro	1) 79110	2	8	UJJAIN	24/02/2016	435	
Micro	1) 10611	2	5	DEWAS	24/02/2016	402	
Micro	1) 10504	2	1	RAJGHAR	24/02/2016	422	
Micro	1) 79110	2	5	UJJAIN	24/02/2016	435	

2.3 Derived and Aggregated Datasets

To support efficient analysis and visualization, the raw dataset was transformed into multiple derived datasets through aggregation and feature engineering. These derived datasets include:

- State-level MSME distribution summaries
- Social inclusion profiles by gender and social category
- Employment and investment aggregates by state and district
- Sector-wise manufacturing versus services composition
- Composite MSME development scores for benchmarking

These datasets reduce computational overhead during visualization and allow focused analysis of specific dimensions without repeated processing of raw records.

2.4 Data Quality and Assumptions

Several data quality measures were applied to ensure analytical reliability:

- Duplicate enterprise records were identified and removed
- Inconsistent categorical labels were standardized across the dataset
- Missing or incomplete values were handled using aggregation-level imputation or exclusion
- Extreme outliers were managed through normalization and comparative analysis

All Aadhaar-related identifiers remained anonymized, and no personally identifiable information was accessed or exposed during the analysis.

The dataset is assumed to be representative of registered MSMEs and does not include unregistered or informal enterprises, which is acknowledged as a limitation in later sections.

3. Methodology

3.1 Data Cleaning and Preprocessing

The methodology begins with a structured data cleaning and preprocessing pipeline to ensure consistency, accuracy, and analytical validity. Since the dataset is sourced from administrative records, particular attention was given to standardization and error handling.

Key preprocessing steps included:

- Standardizing column names and categorical labels to maintain uniformity across records
- Normalizing state and district names to avoid duplication during geographic aggregation
- Extracting valid industry classification information from NIC code fields

- Removing duplicate or incomplete enterprise records that could distort aggregate metrics

These steps ensured that the dataset was structurally consistent and suitable for multi-level analysis at state and district scales.

3.2 Feature Engineering

To enable deeper analytical insights, several derived features were constructed from the cleaned data:

Social Inclusion Indicators

Gender and social category attributes were transformed into percentage-based indicators at state and district levels. These metrics were used to assess inclusivity in MSME ownership and participation.

Employment Efficiency Metrics

Employment efficiency was calculated by relating total employment generated to total investment. This feature highlights regions where MSMEs generate higher employment with relatively lower capital input.

Sectoral Composition Measures

NIC codes were mapped to high-level industry categories to distinguish between manufacturing and services sectors. Counts of unique industry categories were used as a proxy for industry diversity and economic resilience.

Feature engineering allowed raw administrative attributes to be converted into interpretable indicators aligned with societal and economic objectives.

3.3 Composite MSME Development Score

To facilitate comparative benchmarking across states, a composite MSME development score was designed using four core dimensions:

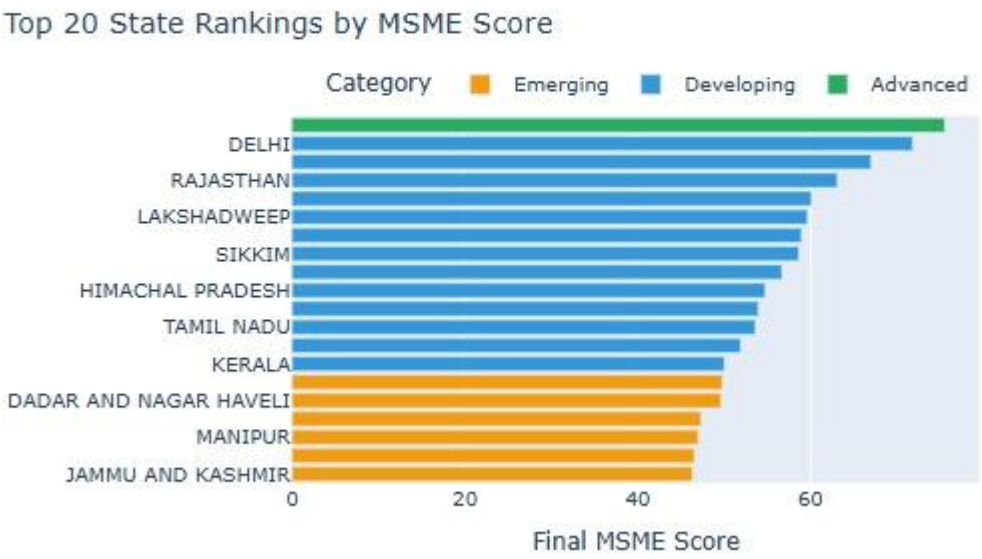
- Scale of MSME activity
- Social inclusion
- Employment generation

- Industry diversity

Each dimension was normalized using min–max scaling to ensure comparability across different measurement units. Equal weightage was assigned to all dimensions to maintain balance and avoid bias toward any single factor.

The resulting composite score provides an objective measure of MSME ecosystem development and enables classification of states into relative development categories for policy analysis.

Development Scorecard

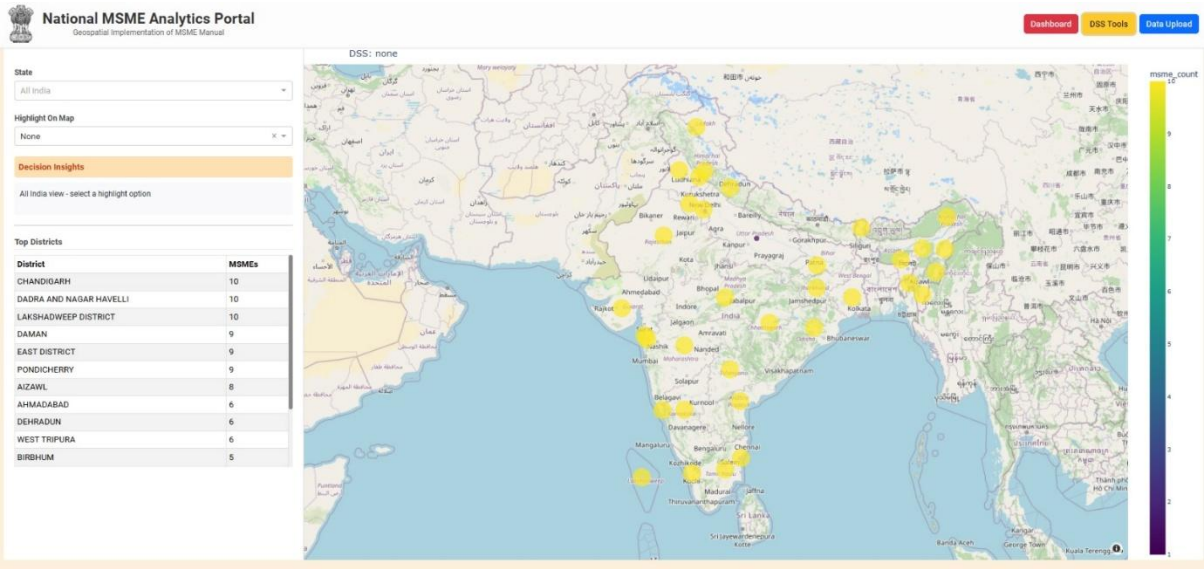


3.4 Analytical Framework

The study employs a layered analytical framework:

- **Univariate analysis** to examine distributions and central tendencies of individual variables
- **Bivariate analysis** to explore relationships between pairs of variables such as investment and employment
- **Trivariate analysis** to study interactions among geography, sector, and employment outcomes

This framework ensures both depth and contextual relevance, allowing insights to be interpreted within broader societal and regional contexts.



4. Data Analysis and Visualisation

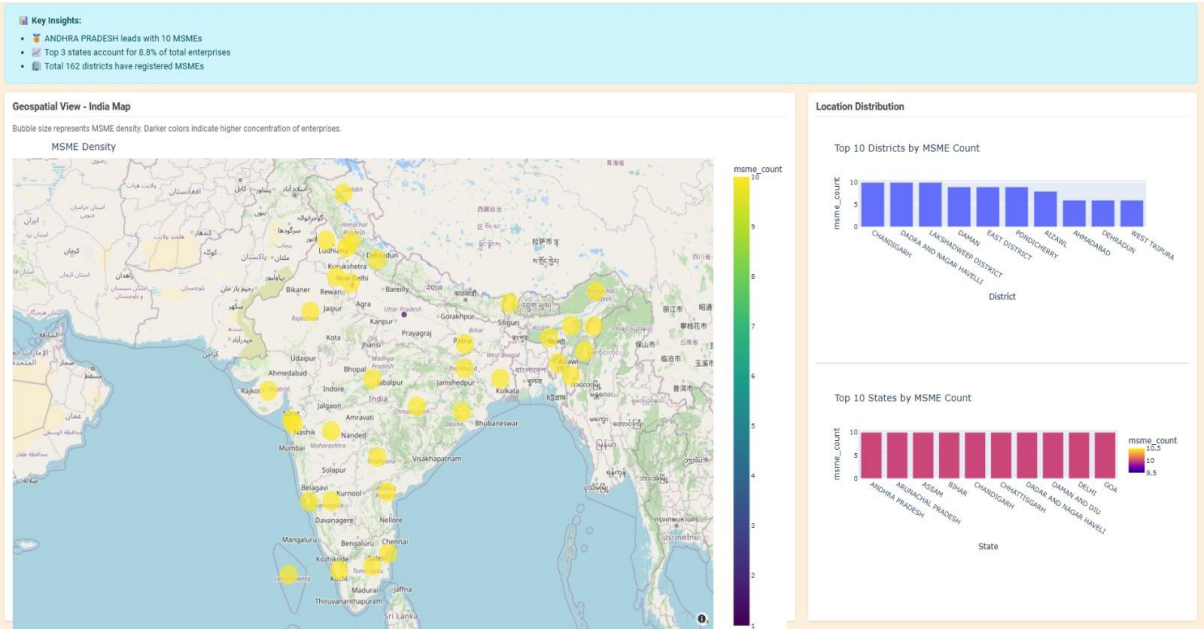
4.1 Overview of Analysis

The analysis focuses on extracting meaningful societal and economic insights from Aadhaar-linked MSME enrolment data using univariate, bivariate, and trivariate techniques. Findings are interpreted at state and district levels to highlight regional disparities and development patterns.

4.2 Key Insights

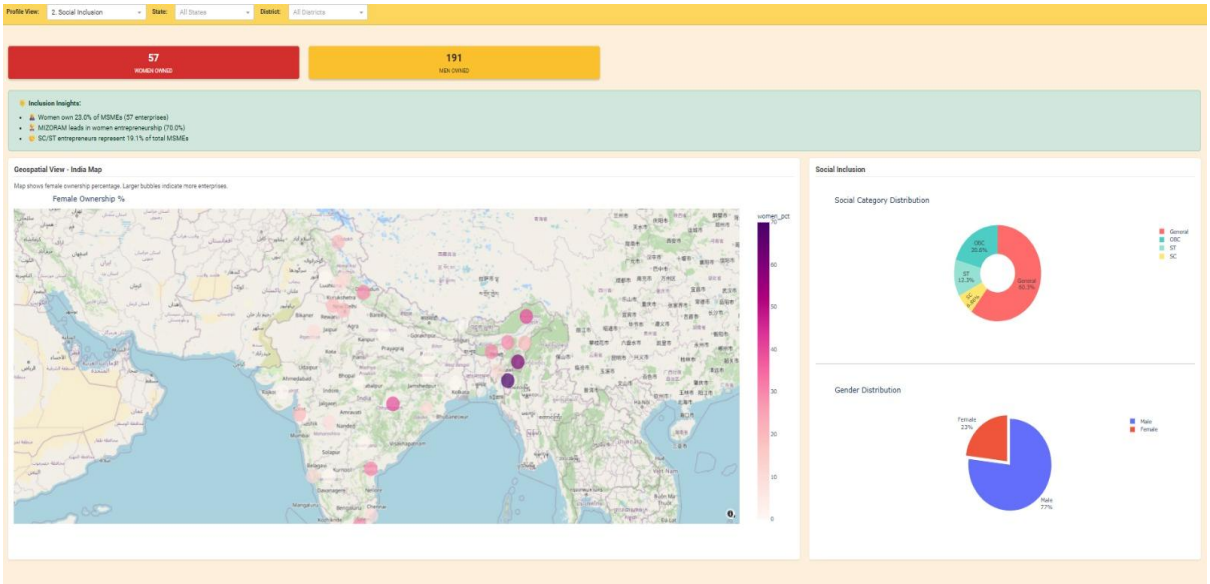
Geographic Distribution

MSME registrations are unevenly distributed across states, with a small number of regions accounting for a large share of total enterprises. This indicates concentration of infrastructure and economic activity.



Social Inclusion

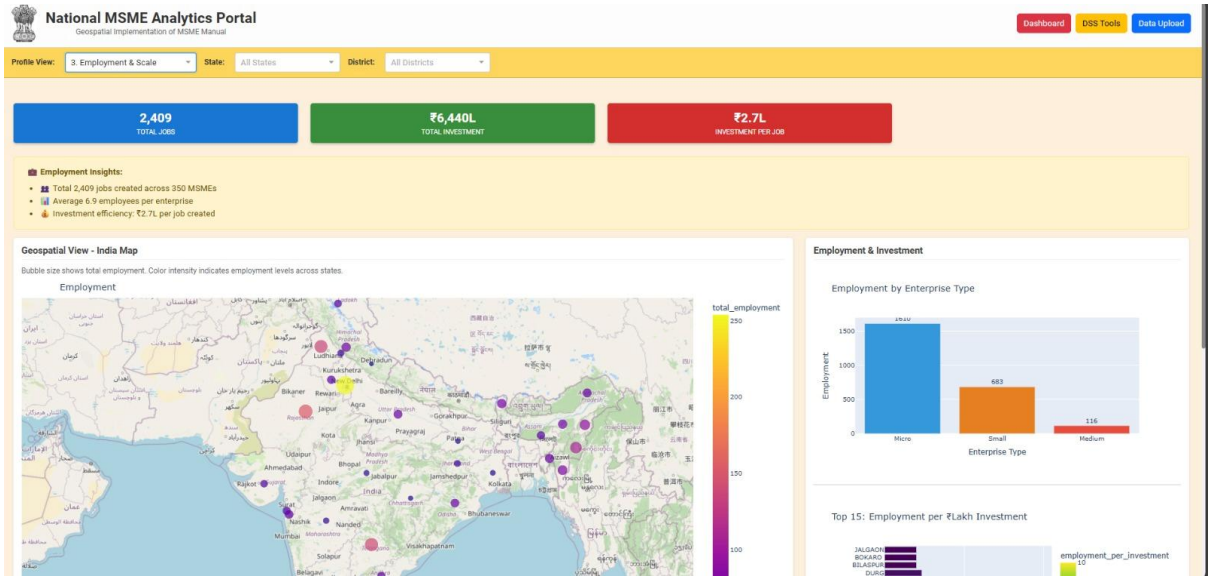
Female participation in MSME ownership remains limited, with significant variation across states. Certain regions demonstrate stronger inclusion outcomes, suggesting the influence of local socio-economic factors.

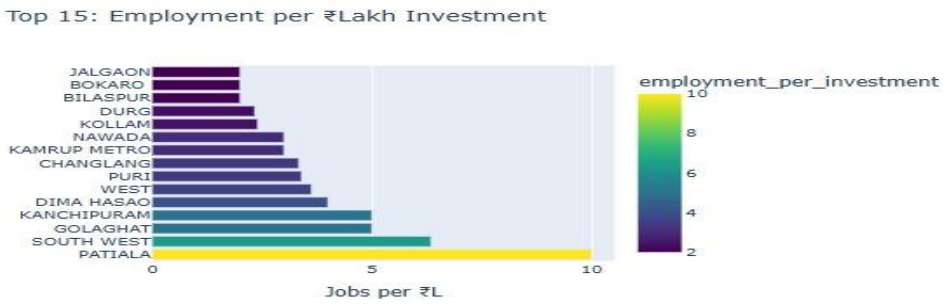


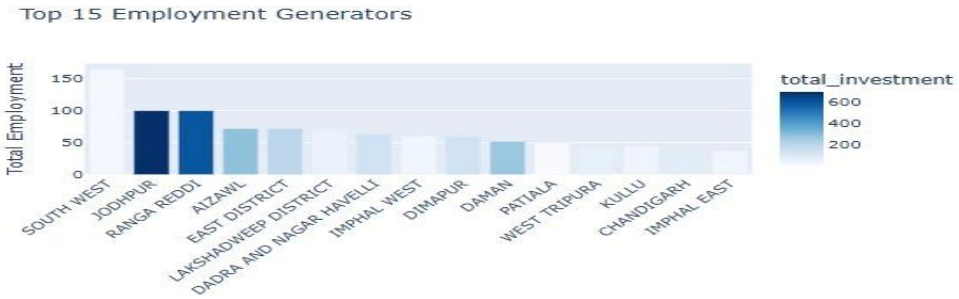
Submitted By ArthNiti Analytics Team

Employment and Investment Relationship

A positive relationship exists between investment and employment; however, several regions exhibit high employment generation even with lower investment, highlighting labor-intensive enterprise models.



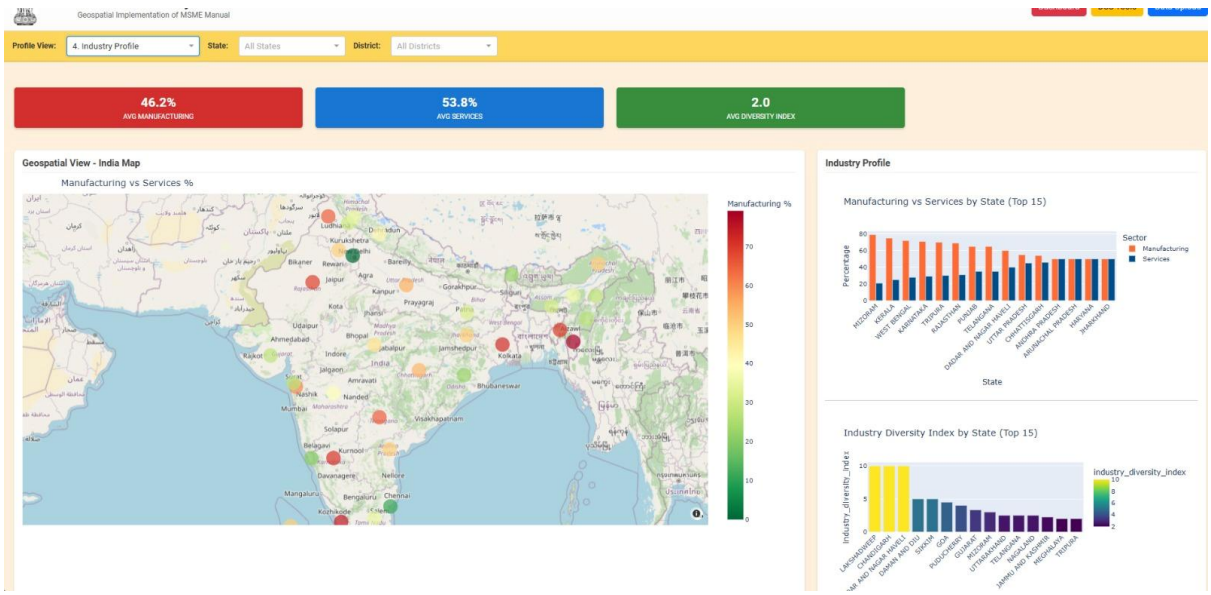




Submitted By ArthNiti Analytics Team

Sectoral Composition

Service-sector MSMEs dominate in many states, while manufacturing remains concentrated in a few regions, raising concerns around sectoral balance and resilience.



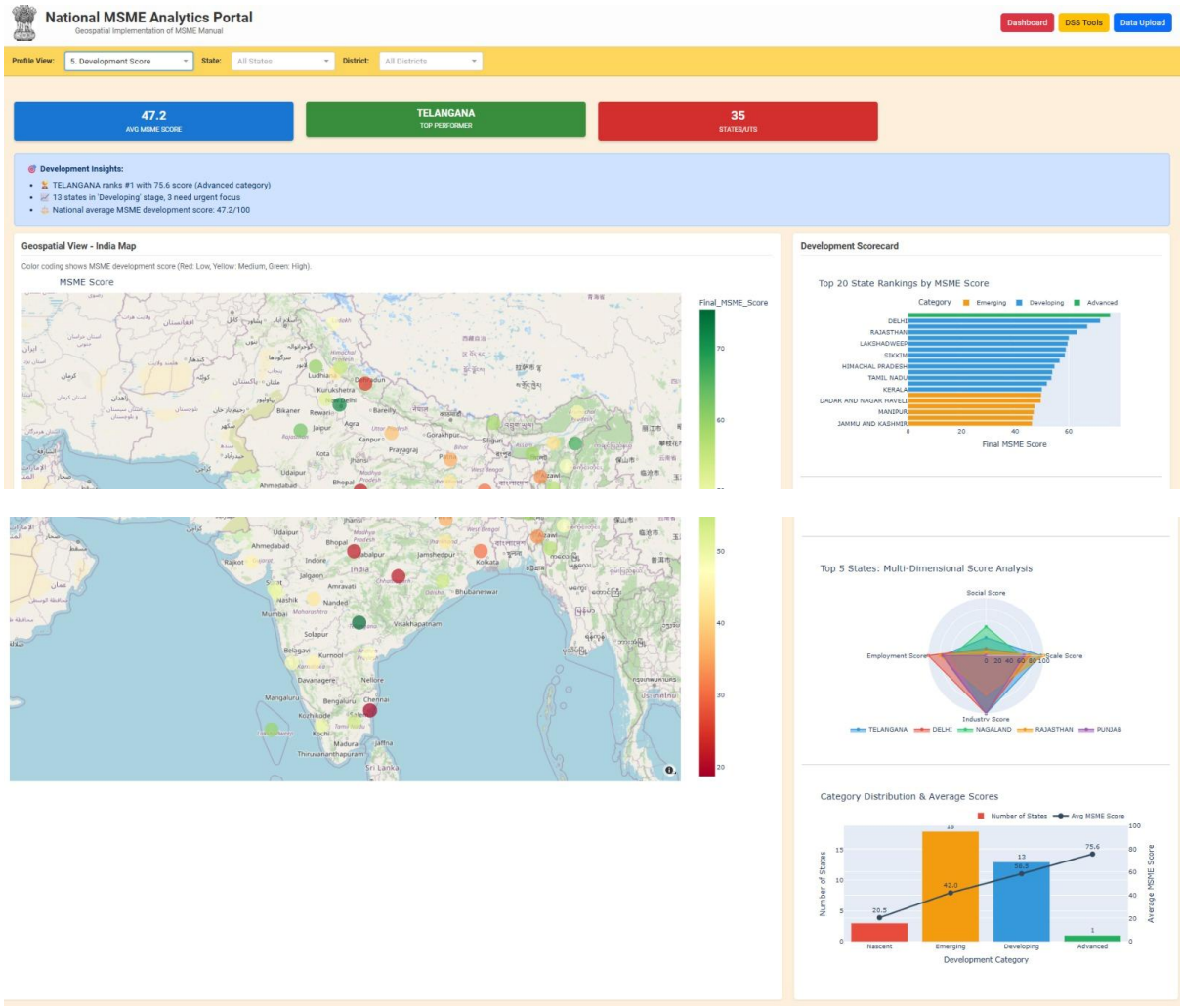
Development Benchmarking

To enable meaningful comparison across regions, a composite MSME development score was used to benchmark states based on multiple dimensions of enterprise performance and inclusion. The scoring framework integrates indicators related to MSME scale, social inclusion, employment generation, and sectoral diversity, ensuring that no single factor disproportionately influences the final assessment.

The benchmarking results reveal a clear clustering of states into distinct development levels. A small group of states consistently demonstrates higher overall performance, characterized by stronger MSME concentration, better employment outcomes, and relatively balanced sectoral composition. These states form the upper tier of MSME development and can be considered benchmarks for best practices in enterprise ecosystem growth.

In contrast, several states fall into lower development clusters, reflecting challenges such as limited enterprise density, weaker employment generation, lower social inclusion, or sectoral concentration. The presence of such clusters

highlights uneven MSME ecosystem development across the country and underscores the need for differentiated, region-specific policy interventions rather than uniform nationwide strategies.



4.3 Visualisation Strategy

Insights are communicated through clear and purpose-driven visualisations, including geospatial maps, bar charts, and comparative plots. Each visual is designed to directly support analytical findings and enable intuitive interpretation for administrative users.

5. Impact and Practical Applicability

The findings of this study are practically applicable in multiple contexts:

- **Policy Design**: Identification of low-performing regions allows targeted policy support instead of nationwide uniform schemes
- **Program Monitoring**: Changes in inclusion and employment indicators can be tracked over time to evaluate policy impact
- **Infrastructure Planning**: Geographic concentration patterns help prioritize industrial and support infrastructure development
- **Financial Inclusion**: Employment efficiency metrics highlight regions suitable for credit and funding interventions

The analytical framework is scalable and can be updated as new Aadhaar-linked MSME data becomes available.

Submitted By ArthNiti Analytics Team

6. Conclusion

This project demonstrates how Aadhaar-linked MSME enrolment data can be transformed from a routine administrative dataset into a powerful analytical resource for understanding societal and economic trends. By combining structured data processing, multi-level analysis, and clear visualisation, the study moves beyond descriptive reporting to generate actionable insights.

The analysis highlights significant regional disparities in MSME distribution, persistent gaps in social inclusion, variations in employment generation efficiency, and imbalances in sectoral composition. The introduction of a composite MSME development score provides a transparent and objective framework for benchmarking states and identifying priority areas for intervention.