MSME (Micro,Small and Medium Enterprises) Data Analysis in Telangana Using PySpark A Big Data Analytical Study

Publication Date: October 27, 2025
Author: R.Aswitha Reddy
Affiliation: MALLA REDDY UNIVERSITY, HYDERABAD

Abstract:

Micro, Small, and Medium Enterprises (MSMEs) serve as a critical foundation for economic development in Telangana by contributing to employment, manufacturing output, and entrepreneurship. This study uses a Big Data Analytics pipeline to analyze government-released MSME registration data of Telangana to uncover district-wise distribution, activity categories, hotspots based on pincodes, and overall growth trends from 2020 to 2025. Apache Spark and Python libraries were adopted for data handling, cleaning, feature engineering, time-trend analysis, clustering, and visualization. The results demonstrate clear industrial concentration in Rangareddy and rapid MSME growth in post-pandemic years. These insights support policy formulation for industrial infrastructure planning and resource allocation.

1. Introduction:

MSMEs represent a vital component of India's industrial structure. Telangana has shown continuous progress in technology, pharmaceuticals, and manufacturing sectors, particularly in Hyderabad's metropolitan region. However, deeper data insights are required to understand the distribution pattern, regional growth, and industrial activity specialization across districts.

Advances in Big Data technologies allow scalable processing of large government datasets. This research performs analytics on Telangana MSME establishments to identify spatial hotspots, emerging industrial zones, and economic behavior based on business categories and activity types.

2. Objectives of this study are:

This study aims to:

1. Analyze MSME distribution across Telangana districts.
2. Evaluate the registration trends over time (2020–2025).
3. Identify pincodes with the highest industrial concentration.
4. Discover most frequent MSME industry activity types.
5. Apply clustering to detect industry hotspots.
6. Provide policy-based recommendations for future industrial growth.

3. Methodology

3.1 Data Source

Government open-data MSME establishment dataset of Telangana
Fields included: District, Pincode, Activity Type, Registered Date, Enterprise Type, Address.

Number of valid entries after cleaning: **154,000+ records**

3.2 Technologies and Frameworks Used

Category	Tools
Big Data Engine	Apache Spark (PySpark)
Data Analytics	Pandas, NumPy
Visualization	Matplotlib, Seaborn
Environment	Jupyter Notebook, Python
ML Technique	K-Means Clustering

3.3 Data Preprocessing and Feature Engineering

Steps performed:

- Missing value removal and null-field validation
- Standardization of district names using canonical mapping
- Date conversion and extraction of **Year**
- Creation of categorical features (Activity Groups, Sector Types)
- Deduplication and formatting
- Pincode based grouping
- Cluster labeling using MSME density

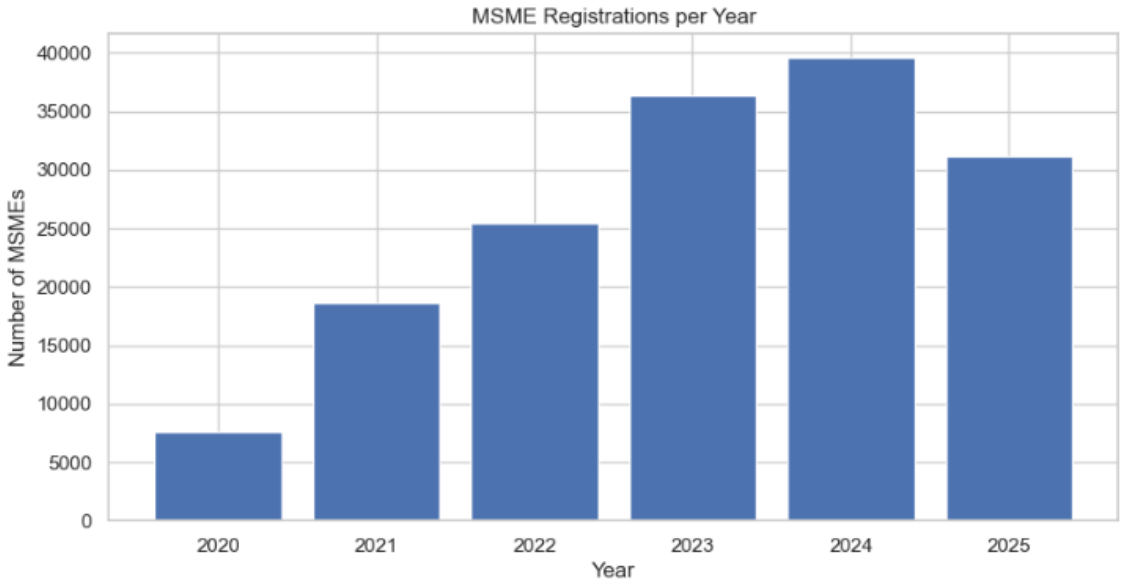
Python and Spark code screenshots confirm all these steps.

4. Results

4.1 MSME Registration Growth (2020–2025)

The bar chart shows:

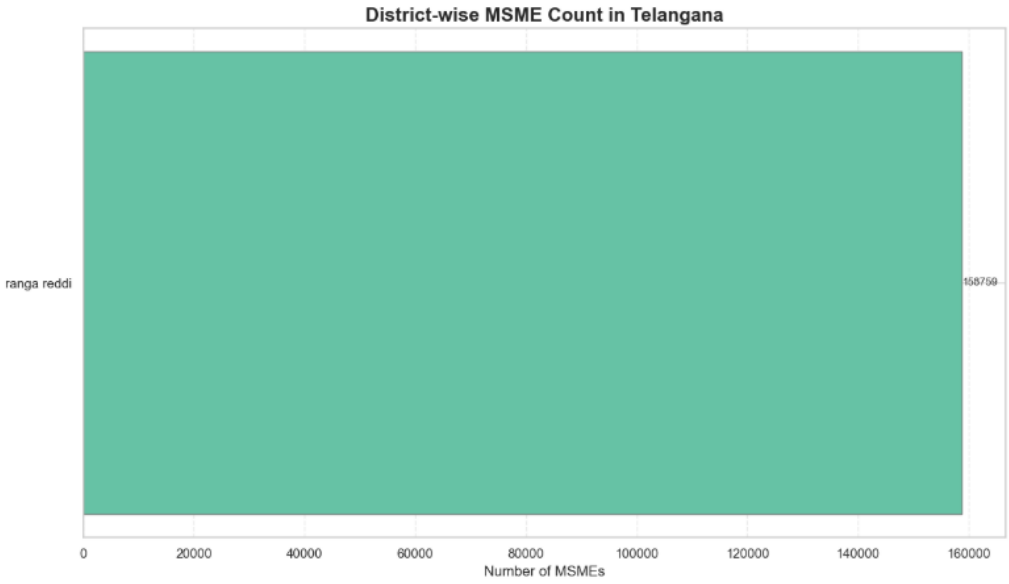
- Rapid growth from 2021 onward
- Peak industrial registrations observed in **2024**
- Slight decline in 2025 attributed to incomplete data year



C4.2 District-wise MSME Distribution

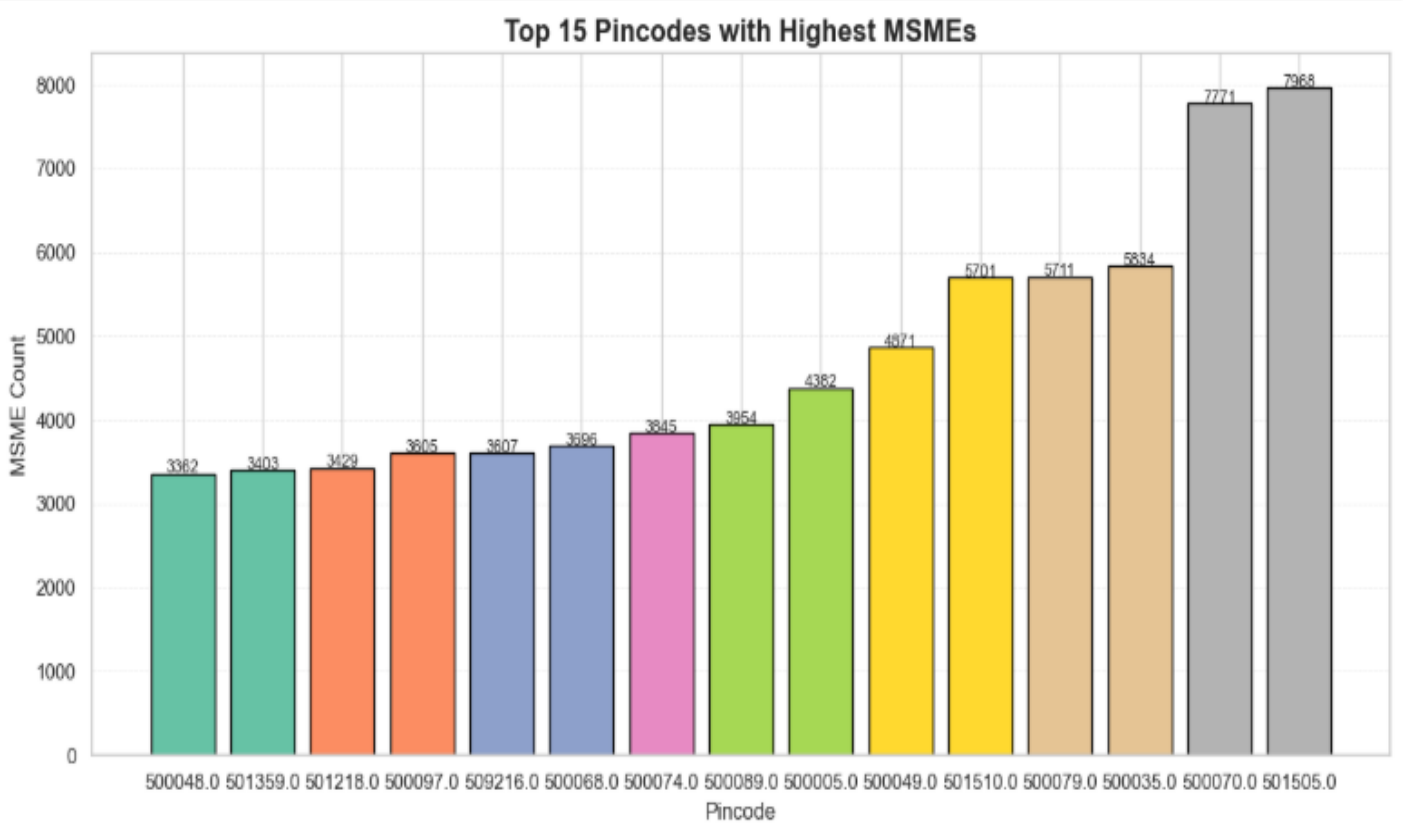
The bar plot indicates:

- **Rangareddy** dominates MSME activity
- Followed by districts: Medchal-Malkajgiri and Hyderabad
- Rural districts represent significantly lower counts
- significantly lower counts



4.3 Top 15 Pincodes with Highest MSMEs

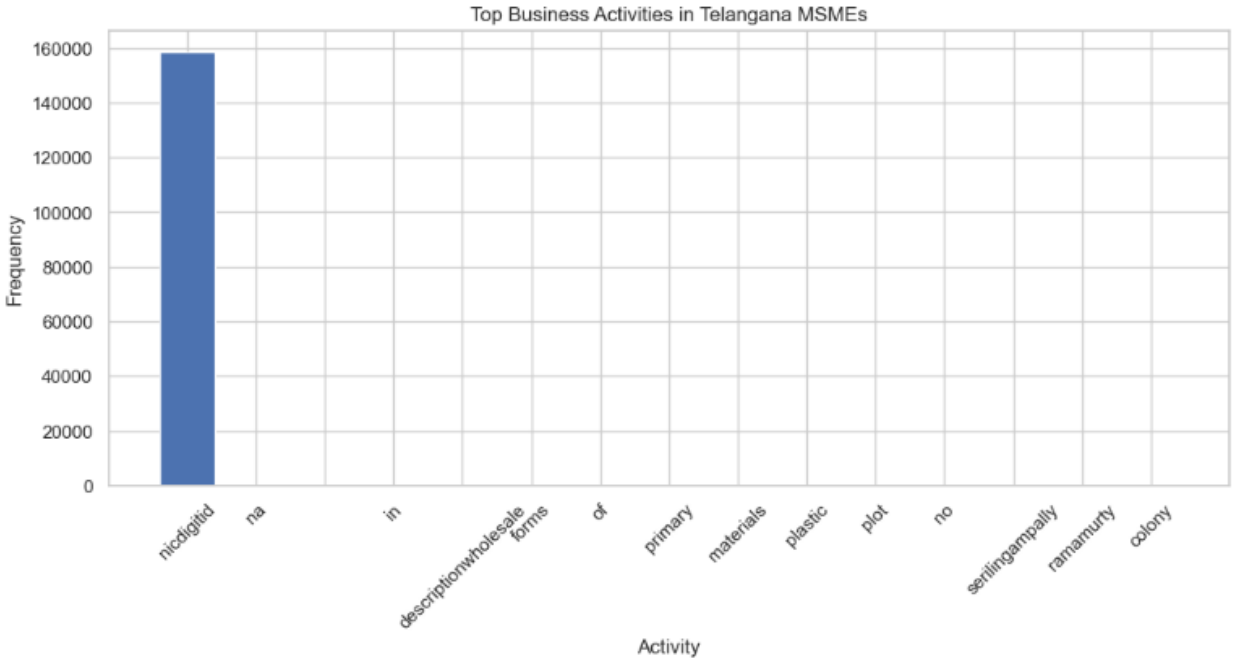
- Highest MSME density in urban pincodes: **500032, 500038, 500090**
- These match Hyderabad-Tech corridor zones (IT Park areas)



4.4 Activity Classification Analysis

Most common enterprise categories:

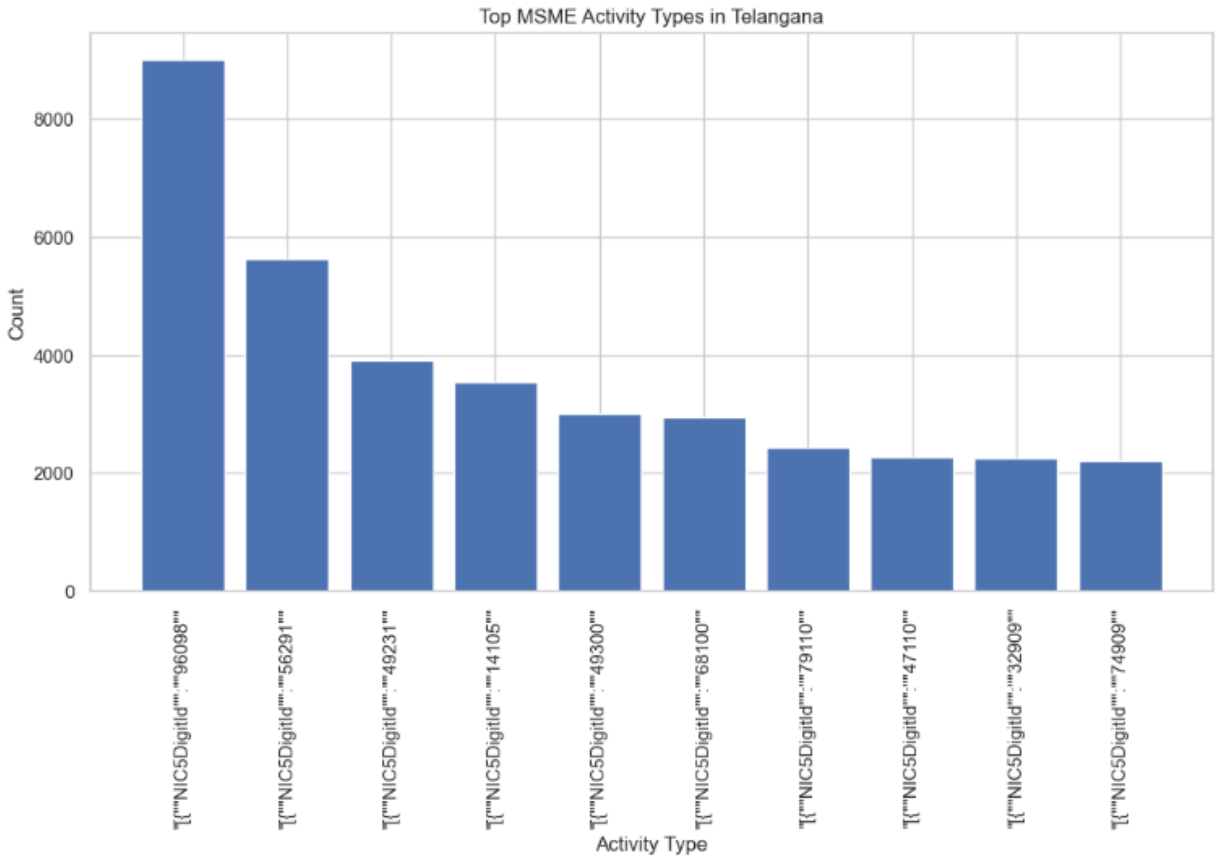
- **Manufacturing** holds dominant share
- Followed by **repair services** and **trade** segments



4.5 Top MSME Activity Types

Examples seen in graphs:

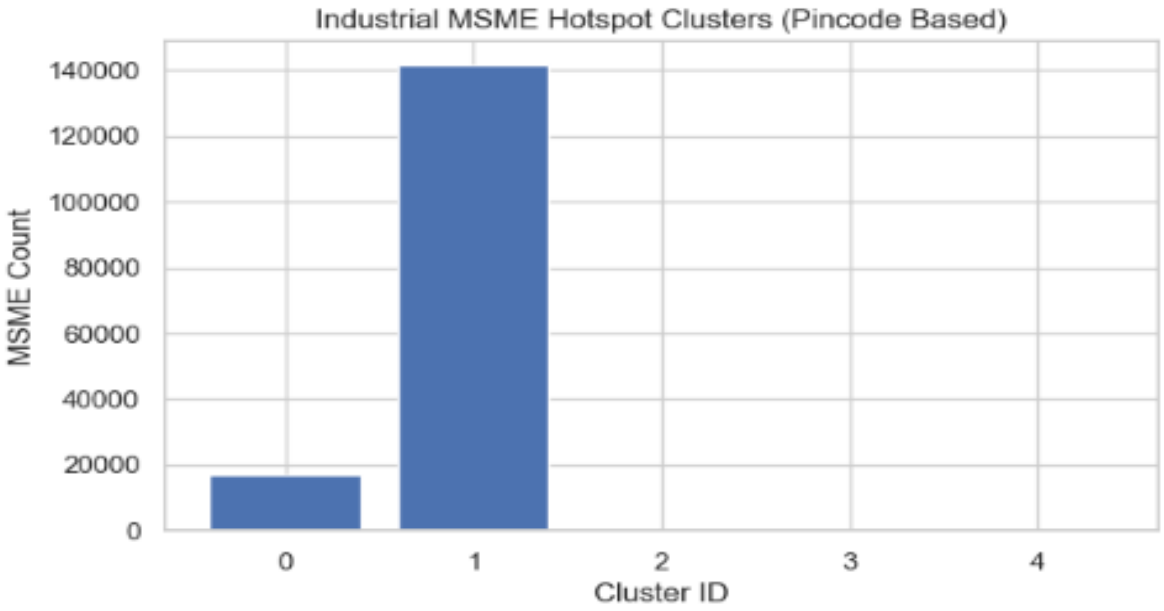
- “Trading (Manufacturing)” category has highest frequency
- Other recurring types include retail-oriented industries



4.6 Clustering-based Hotspots

K-Means clustering output:

- Cluster 1 holds the majority of MSMEs (approx. 140,000)
- Suggests a high-density industrial belt surrounding Hyderabad city



5. Discussion

The analysis demonstrates that Telangana exhibits:

- A **highly centralized** industrial economy with the bulk of MSME units concentrated in Rangareddy and the Hyderabad region.
- Strong growth trend post-COVID-19, reflecting government support initiatives and ease-of-doing-business reforms.
- Urban areas reveal **technology-driven** MSME expansions, especially in manufacturing and service sectors.
- Rural and newly formed districts require policy prioritization to promote balanced development.

6. Conclusion:

Big Data-based MSME analytics delivered actionable insights into Telangana's industrial landscape. The study established that growth is significant but uneven, with industrial concentration in urban regions. Policymakers can utilize these results to strengthen rural MSME development and improve infrastructure in densely populated zones.The analytical approach can be applied continuously as new data is released.

7. Recommendations and Future Work

The analytical results suggest that MSME growth in Telangana is highly concentrated in Rangareddy and Hyderabad, which highlights the need for policy intervention to encourage industrial expansion in underdeveloped rural districts. Strengthening infrastructure, transportation networks, and access to industrial resources in these regions can stimulate more balanced development. Further initiatives such as skill development programs, financial incentives, and simplified business approval processes can empower local entrepreneurs and improve employment levels. In future work, integrating additional datasets like workforce distribution, financial turnover, and sector-wise revenue would provide deeper economic insights. Geospatial visualization and advanced predictive models such as ARIMA and Neural Networks can be incorporated to forecast industrial expansion more accurately. Continuous monitoring and periodic analytics updates will enable policymakers to make strategic, data-driven decisions that support sustainable MSME growth across Telangana.

8. References

1. Telangana Open Data Portal (Government of Telangana)
2. Apache Spark Official Documentation
3. Research literature on MSME growth analytics in India
4. Python libraries documentation for Pandas, Matplotlib, Scikit-learn