Spark Mini Project Report

Prepared By: **Ch. Pavan Sai**

Course: **Big Data Analytics**

Date: **October 2025**

Dataset Description (pincode dataset.csv)

- Columns Identified:

  • officename → Name of the post office

  • pincode → 6-digit postal code identifying a region

  • officetype → Type of post office (Head, Sub, Branch)

  • deliverystatus → Whether delivery services are available

  • district / statename → Administrative region information

  • latitude / longitude → Coordinates for geospatial analysis

■ Observations from Executed Cells

1. Data Loading & Cleaning: Handled missing latitude and longitude values; standardized district and state names; converted coordinates to numeric format.

2. State & District-Level Analysis: Found that large states such as Uttar Pradesh, Maharashtra, and Bihar have the highest number of post offices, while smaller states and UTs have fewer.

3. Office-Type Distribution: Majority of offices are Branch Offices followed by Sub and Head Offices.

4. Delivery Status: Most offices provide delivery services, indicating strong operational coverage nationwide.

5. Interactive Mapping (Folium): Created an India-wide map showing all postal locations using clustering for performance.

6. Nearby Post Office Finder: Developed a module to locate post offices within 15 km of any given location (based on latitude & longitude).

7. Suitable Location for New Offices: Introduced a Saturation Score = Total Offices / Unique Pincodes to identify underserved regions where offices are far apart (> 5 km).

8. Postal Network Optimization: Combined analytical scoring with DBSCAN clustering to detect isolated post offices, suggesting priority zones for expansion.

9. Population vs Office Count: Compared each state’s population to post office count to assess service adequacy per capita.

■ Plots & Visualizations Observed

  • State-wise and district-wise office count (bar chart)

  • Office type and delivery status (pie charts)

  • Nearby post office locator (interactive map)

• Postal saturation map (green = underserved regions)
• Isolated post offices visualization (purple markers)
• State population vs postal network comparison (bar graph)

■ Key Insights

- Branch Offices form the backbone of the Indian postal network.
- Several rural districts remain underserved based on low saturation scores.
- Geospatial clustering effectively identifies isolated offices beyond 5 km.
- The integration of population data highlights postal service inequality between states.

■ Recommendations

1. Increase postal infrastructure in low-saturation and rural districts.
2. Use spatial clustering models to optimize service placement.
3. Integrate real-time population and demographic data for better planning.

■ Conclusion

The project successfully demonstrates how PySpark and geospatial analytics can be applied to real-world datasets for postal network optimization. The findings support data-driven decision-making for improving coverage, accessibility, and planning of postal services across India.