All India Pincode Data Analysis Report

Dataset Description

The dataset used in this analysis is the "All India Pincode Directory," which provides exhaustive information about postal offices across India. The raw CSV file contains 165,627 records corresponding to various post offices, with columns such as: circlename, regionname, divisionname, officename, pincode, officetype, delivery, district, statename, latitude and longitude (geolocation coordinates).
Total Rows: 165,627
Unique Pincodes: 19,586
Unique States: 37
Unique Districts: 750

Operations Performed

Data Loading & Schema Inspection
- Loaded CSV into a Spark DataFrame with schema inference.
- Inspected column types: All string except pincode (integer) and coordinate columns (string, later cast to numeric for analysis).

Exploratory Analysis
- Counted overall records and unique values for pincodes, states, districts, office types, delivery status, and detected schema validity.
- Displayed the first 10 records for data familiarity.

Key Aggregations & Grouping
- State-Level Aggregation: Counted pincodes per state (e.g., Tamil Nadu has the most: 2049).
- District-Level Aggregation: Counted pincodes and offices per district.
- Region and Circle Aggregation: Examined regions under circles, divisions under regions, and districts under circles.
- Office Type Distribution: Aggregated HO/PO/BO counts per state.

Duplicate & Distribution Analysis
- Identified pincodes shared by multiple offices (duplicates).
- Plotted distribution of office counts per pincode.

Location Validity Checks
- Checked for missing/null/zero latitude and longitude entries, and listed offices/states with invalid location data.

Data Visualization
- Plotted stacked bar charts for office type breakdowns by state.
- Bar charts illustrating invalid/missing coordinates by state.
- Histograms showing distribution of offices per pin code.

Key Insights

- Tamil Nadu has the highest number of unique pincodes, followed by Uttar Pradesh and Maharashtra.
- The top 5 districts with most pincodes: "NA", Thrissur, Pune, 24 Paraganas North, Ernakulam.
- The majority of offices are Delivery offices (157,901 out of 165,627).
- Typical office types are HO (Head Office), PO (Post Office), and BO (Branch Office); BO is most common.
- Some pincodes are shared among multiple offices, especially in urban metros (Delhi, Mumbai).
- North Eastern Circle covers the largest number of districts (85).
- Andhra Pradesh (Y.S.R. District) has the highest number of offices in a single district-state combination (819 offices).
- There are offices with missing or invalid coordinates (latitude/longitude), impacting geospatial tasks.
- No missing/null values in core columns, except for some invalid coordinates.
- Most pincodes are unique to a single office, but urban pincodes often span several.

Recommendations

- **Data Quality**: Address missing or zero geolocation values, especially for offices in Bihar and areas with frequent nulls, for better mapping and geospatial analysis.
- **Postal Planning:** Concentrate infrastructure upgrades in districts and states with high office density and delivery volume, notably in Tamil Nadu and Andhra Pradesh.
- **Duplicate Handling:** Investigate urban pincodes with multiple offices to optimize delivery strategies and reduce confusion.
- **Further Analysis:** Include population, service requests, and delivery success rates per pincode for richer insights.
- **Spatial Mapping:** Integrate coordinate corrections to enable proper visualizations and route planning.
- **Office Type Optimization:** Explore why some states have disproportionate BO/PO/HO distributions to rationalize administrative workloads.

Example Snippet

Column Name	Description	Sample Value
circlename	Postal Circle name	Bihar Circle
regionname	Region name	East Region
divisionname	Division name	Nalanda Division
officename	Office name	Tinalodipur BO
pincode	Postal code	801305
officetype	HO/PO/BO	BO
delivery	Delivery/Non Delivery	Delivery
district	Name of the district	PATNA
statename	Name of the state	BIHAR
latitude	Geolocation latitude	27.88
longitude	Geolocation longitude	79.91