UIDAI DATA HACKTHON 2026

DATASET: api_data_aadhar_biometric

Team id : UIDAI_9366

INTRODUCTION:

This dataset contains aggregated information on biometric updates (modalities such as fingerprints, iris, and face). It reflects the periodic revalidation or correction of biometric details, especially for children transitioning into adulthood.

STEPS INVOLVED:

1. **Data cleaning and preparation**
2. **Exploratory Data Analysis**
3. **Insights / key Findings**
4. **Recommendations**

DATA CLEANING AND PREPARATION

- **Initial Data Loading**: Four CSV files were successfully loaded and concatenated into a single DataFrame, initially comprising 1,861,108 entries across 6 columns.
- **Missing Values**: No missing values were detected in any of the columns, indicating a complete dataset from the start.
- **Duplicate Removal**: A total of 94,896 duplicate rows were identified and removed, reducing the dataset size from 1,861,108 to 1,766,212 unique entries.
- **Data Type Standardization**: The 'date' column was successfully converted from an object type to datetime64[ns] for accurate temporal analysis.
- **Date Column Splitting**: The 'date' column was split into three new integer columns: 'year', 'month', and 'day', facilitating granular temporal analysis.
- **Error Correction**:
    - The 'state' column was extensively cleaned, standardizing various misspellings and casing inconsistencies (e.g., 'Orissa' to 'Odisha', multiple 'West Bengal' variations to 'West Bengal') and consolidating entries for merged territories.
    - The 'district' column was cleaned by removing special characters (like '*' and '?'), standardizing casing to title case, and replacing invalid entries with None. This reduced the number of unique districts from 974 to 895.

EXPLORATORY DATA ANALYSIS:

- **Explored the Dataset:** Performed analysis on data with info, describe, shape
- **Outlier Handling**: Statistical outliers were identified in the 'bio_age_5_17' (201,325 entries) and 'bio_age_17_' (210,205 entries) columns using IQR. However, these were retained, as they were determined to be valid extreme observations of biometric scan counts rather than data errors, crucial for maintaining data integrity.
- **Final Dataset State**: The cleaned DataFrame consists of 1,766,212 rows and 9 columns, with appropriate data types assigned for each column ('date' as datetime64[ns], 'state' and 'district' , 'pincode', 'bio_age_5_17', 'bio_age_17_' as int64, and 'year', 'month', 'day' as int32).

DATA ANALYSIS KEY FINDINGS:

- The bio_age_5_17 and bio_age_17_ columns, representing biometric activity for different age groups, both exhibited highly right-skewed distributions. A substantial number of outliers were identified: 201,325 for bio_age_5_17 and 210,205 for bio_age_17_, indicating instances of unusually high activity.
- Analysis of biometric activity over time (from March to November 2025) revealed fluctuating monthly trends. Both bio_age_5_17 and bio_age_17_ showed similar patterns, peaking around July-August and declining towards the end of the year.
- Geographically, Uttar Pradesh and Maharashtra emerged as the states with the highest total biometric activity, with 9,367,083 and 9,020,710 total activities, respectively. Within these top states, biometric activity is concentrated in specific districts, such as Pune in Maharashtra and Ghaziabad in Uttar Pradesh.
- A strong positive linear relationship was observed between bio_age_5_17 and bio_age_17_ (correlation coefficient of 0.79), indicating that areas with higher biometric activity in one age group tend to have higher activity in the other.
- Both bio_age_5_17 and bio_age_17_ showed moderate negative correlations with the month column (approximately -0.37 to -0.39) and weak negative correlations with the day column (approximately -0.18 to -0.19), suggesting higher biometric activity in the earlier parts of the recorded months and days.

RECOMMENDATIOS:

Based on these findings, here are some recommendations:

1. **Strategic Resource Allocation:** Prioritize resource allocation (e.g., more enrollment centers, staff, mobile units) to states and districts with persistently low biometric activity. This includes areas like Salumbar (Rajasthan), Bandipur (Jammu and Kashmir), and Nicobars (Andaman and Nicobar Islands), to improve accessibility and awareness.

2. **Investigate Low-Activity Areas:** Conduct localized surveys or qualitative studies in districts with extremely low activity to understand the root causes. This could be due to low population density, geographical barriers, lack of awareness, or specific socio-economic factors.

3. **Learn from High-Activity Regions:** Analyze the factors contributing to high biometric activity in states like Uttar Pradesh and Maharashtra, and particularly in districts like Pune. Identify best practices, successful outreach strategies, and efficient operational models that could be replicated or adapted in other regions.

4. **Targeted Outreach and Awareness Campaigns**: For low-activity areas, implement targeted awareness campaigns, possibly in local languages, to educate residents about the importance and benefits of biometric enrollment.

5. **Data Quality and Verification:** For districts showing exceptionally low numbers (e.g., 1-8 activities), it would be prudent to perform a data quality check to ensure these figures accurately reflect the ground reality and are not due to reporting gaps or errors.

6. **Policy Development for Regional Disparities:** Develop policies that specifically address regional disparities in biometric activity, ensuring equitable access to services across all states and districts.