Demographic Trends Analysis Using PySpark

■ Dataset Description

The dataset, sourced from Demographic_update_data_March-July.csv, contains demographic information aggregated at the district level across various states in India. The data spans the period from March to July 2025 and captures daily records.

The dataset includes the following columns:
- **Date**: The date of data entry (DD-MM-YYYY).
- **State**: The Indian state where the data was recorded.
- **District**: The district within the state.
- **Pincode**: The postal code of the area.
- **Demo_age_5_17**: Population aged 5–17.
- **Demo_age_17+**: Population aged 17 and above.

■ Code Cells Description

1. Initial Setup and Data Loading

The Spark environment is initialized with a SparkSession to load the CSV into a Spark DataFrame. Initial schema inspection reveals that all columns, even numeric ones, are read as strings.

2. Data Cleaning and Preparation

Missing values in numeric columns (Pincode, Demo_age_5_17, Demo_age_17+) are replaced with 0. The Date column is reformatted to a proper date type, and Month and Year columns are derived for time-based analysis.

3. Data Aggregation and Feature Engineering

The data is grouped by State and Month to compute totals for youth and adult populations. Derived metrics include:
- **total_population** = youth + adult
- **pct_age_5_17** = percentage of youth
- **mon_growth_pct** = month-over-month growth percentage using a Spark window function.

4. Filtering for Analysis

The dataset is filtered to the latest month (July 2025) for focused analysis on top states and districts by population metrics.

■ Visualizations and Insights

Fig 1: Top 10 States by Youth Share

Observation: Madhya Pradesh shows the highest youth share (≈20%), followed by Chandigarh and Gujarat. Youth proportions vary significantly across states.

Fig 2: Month-over-Month Population Growth
Observation: A sharp decline across major states between March and April 2025, followed by mixed recovery trends. Maharashtra and Uttar Pradesh stabilize, while Karnataka continues to decline.

Fig 3: Population Composition for Top 10 States
Observation: Uttar Pradesh leads in total population, with adults forming the majority in all states.

Fig 4: Population Distribution in Madhya Pradesh
Observation: Adults make up 80.6%, youth 19.4%. Nearly one-fifth of Madhya Pradesh's population is youth.

Fig 5: Top 15 Districts by Total Population
Observation: Bangalore tops the list, followed by Thane and Pune. Maharashtra and West Bengal dominate the top 15.

Fig 6: Distribution of Youth Population Share by Month
Observation: The median youth share stays consistent (10–12%) across months, with more variability in April and July.

Fig 7: Distribution of Total Population per District
Observation: Most districts have low populations; Uttar Pradesh and Maharashtra show high variance with several outlier districts.

Fig 8: Correlation Matrix of Population Metrics
Observation: Strong positive correlations among population metrics (0.97–1.00). The month-over-month growth rate shows weak correlation with population size.

Conclusion

This PySpark-based demographic analysis from March–July 2025 highlights significant regional differences across India. Uttar Pradesh leads in total population, while Madhya Pradesh has the highest youth proportion. A sharp population drop in April 2025 across major states suggests potential data inconsistencies requiring further investigation. The findings also indicate that a state's population size does not directly predict its short-term growth rate.