📄 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 appears to be recorded on a daily basis for the period of **March to July 2025**.

The dataset includes the following columns:

- **Date**: The date of the data entry (format: DD-MM-YYYY).
- **State**: The Indian state where the data was recorded.
- **District**: The specific district within the state.
- **Pincode**: The postal code of the area.
- **Demo_age_5_17**: The population count for the youth demographic, aged 5 to 17.
- **Demo_age_17+**: The population count for the adult demographic, aged 17 and above.

💻 Code Cells Description

This section provides a step-by-step summary of the data processing and analysis performed in the notebook.

1. Initial Setup and Data Loading

The notebook begins by initializing the Spark environment and creating a SparkSession, which is the entry point for all Spark functionality. It then reads the Demographic_update_data_March-July.csv file into a Spark DataFrame. An initial inspection of the schema reveals that all columns, including numerical ones, are first read as the string data type.

2. Data Cleaning and Preparation

To ensure data quality, the code first addresses missing values. Any **null values** in the numeric columns (Pincode, Demo_age_5_17, Demo_age_17+) are replaced with 0. Following this, the Date column is converted from a string to a proper date format, and new columns for Month and Year are extracted to enable time-based analysis.

3. Data Aggregation and Feature Engineering

The core of the analysis involves aggregating the data. The raw daily data is grouped by State and Month to calculate the total youth and adult populations. From this aggregated data, several new analytical columns are engineered:

- total_population: The sum of youth and adult populations.
- pct_age_5_17: The percentage of the total population that falls into the 5-17 age group.
- mon_growth_pct: The month-over-month percentage growth in total population, calculated using a Spark window function.

4. Filtering for Analysis

Finally, the notebook filters the processed data to focus on specific insights. It identifies the latest available month in the dataset (July 2025) and uses this to create subsets of data for visualizing the top states and districts based on various population metrics.

📊 Visualizations and Insights

This section contains every plot generated by the notebook, each followed by a detailed observation.

Plot 1: Top 10 States by Youth Share

Observation: For the latest month (July 2025), **Madhya Pradesh** has the highest share of its population in the 5-17 age group, at nearly **20%**. The youth share among the top states varies significantly, from over 18% in Chandigarh down to around 13% for states like Telangana and Haryana, indicating different demographic structures across the country.

Plot 2: Month-over-Month Population Growth

Observation: A key trend observed is the **significant negative growth** (a sharp population drop) for all four major states between March and April 2025. After this dip, growth patterns diverge. While Maharashtra and Uttar Pradesh show a more stable and slightly positive trend, Karnataka and Andhra Pradesh continue to decline.

Plot 3: Population Composition for Top 10 States

Observation: **Uttar Pradesh has the largest total population** by a considerable margin compared to the other top states. In all listed states, the adult population (Age 17+) forms the vast majority. The proportion of the youth population (the blue segment) appears relatively consistent across these highly populated states.

Plot 4: Population Distribution in Madhya Pradesh

Observation: In Madhya Pradesh, the adult population (Age 17+) makes up the vast majority at **80.6%**, while the youth population (Age 5-17) accounts for the remaining **19.4%**. This visualization confirms the data from the initial bar chart, highlighting that nearly one-fifth of the state's population is in the youth category.

Plot 5: Top 15 Districts by Total Population

Observation: **Bangalore** district has the highest total population, followed closely by Thane and Pune. The most populous districts are concentrated in a few states, with **Maharashtra** (Thane, Pune, Mumbai Suburban, etc.) and **West Bengal** (North & South 24 Parganas) being heavily represented in the top 15.

Plot 6: Distribution of Youth Population Share by Month

Observation: The **median youth population share** across all states remained relatively stable, hovering around **10-12%** from March to July. However, months 4 (April) and 7 (July) exhibit a **wider spread** and a larger interquartile range, indicating greater variability in the youth population percentage among different states during those months.

Plot 7: Distribution of Total Population per District

Observation: The violin plot shows that most districts in the selected states have a low total population, indicated by the wide base of the violins near the bottom. **Uttar Pradesh** and **Maharashtra** display the greatest variance, with long upper tails showing that they have several outlier districts with significantly higher populations than the typical district.

Plot 8: Correlation Matrix of Population Metrics

Observation: There is a very **strong positive correlation (0.97-1.00)** between the absolute population counts, as expected. Interestingly, the month-over-month growth percentage (mon_growth_pct) shows a **very weak correlation** with all other metrics, suggesting that a state's population size is not a strong indicator of its recent growth rate.

Conclusion:

This analysis of demographic data from March to July 2025 highlights significant regional differences, with Uttar Pradesh leading in total population and Madhya Pradesh in youth share. The most critical finding is a sharp, unexplained population drop across major states in April 2025, which suggests a data anomaly requiring further investigation. The project ultimately reveals that a state's population size is not a reliable predictor of its recent growth rate.