Users Dataset: Detailed Explanation & Insights

The Users Dataset contains demographic and account-related details of customers. The dataset was explored, cleaned, and analyzed using EDA (Exploratory Data Analysis) techniques. Below is a structured summary of the investigation.

Understanding the Data

The dataset includes the following key columns:

- **ID** → Unique identifier for each user.
- **CREATED_DATE** → Date when the user account was created.
- **BIRTH_DATE** → User's date of birth, used to calculate age and generation.
- **STATE** → User's state of residence.
- **LANGUAGE** → Primary language used by the user.
- **GENDER** → User's self-reported gender.

Initial Observations

- The dataset contains a diverse set of users from different states, age groups, and languages.
- Some key fields (BIRTH_DATE, STATE, LANGUAGE, and GENDER) have missing values.
- Duplicate records were detected, which could lead to overrepresentation of certain users in analysis.

Exploratory Data Analysis (EDA)

EDA was conducted to uncover **patterns, distributions, and potential issues** in the dataset. The following visualizations were created:

Missing Values Visualization

- **Bar Chart** :
  - Showed that birth dates, states, languages, and gender fields had missing values.
  - Languages had the highest number of missing values, indicating potential data collection issues.

Age Distribution Analysis

- **Histogram with KDE (Kernel Density Estimation)**:
  - Showed the spread of user ages, helping identify which age groups are most represented.
  - Helped validate missing age data by checking distributions.

Geographic Distribution of Users

- **Bar Chart (Top 10 States)**:
  - Highlighted the states with the highest user counts.
  - Useful for regional segmentation and targeting.

Account Creation Trend

- **Line Chart (User Signups Over Time)**:
  - Showed when most users signed up.
  - Helped identify seasonal trends or spikes due to marketing campaigns.

Assumptions Made

- **Missing Birth Dates & Age Calculation**:
  - Users with missing `BIRTH_DATE` were assigned a default date of `"1970-01-01"`.
  - Users with default birthdates were not included in age-based insights.
- **Missing State & Language**:
  - Unknown states and languages were replaced with `"Unknown"`.
  - Assumed that missing values were due to incomplete user profiles rather than actual missing information.
- **Duplicate User Records**:
  - Considered exact duplicate records (i.e., same `ID` and other details) as erroneous.
  - Assumed that each ID should be unique.

Why Age was left Empty - During EDA, we identified missing values in `BIRTH_DATE`, which prevented proper age calculation. Since `AGE` depends on `BIRTH_DATE`, it was left empty instead of computing incorrect values.

Handling Missing Values & Duplicates

Missing Values

- **Filled missing values** for:
  - `STATE`, `LANGUAGE`, `GENDER` → `"Unknown"`
  - `BIRTH_DATE` → `"1970-01-01"` (default)
  - `AGE` was recalculated from `BIRTH_DATE`

Duplicates

- Removed all duplicate rows based on `ID`.

Key Takeaways

- The dataset contained missing demographic details, especially in `BIRTH_DATE`, `LANGUAGE`, and `STATE`.
- Duplicates were present and successfully removed.
- User distribution was not uniform across states and age groups.
- Certain states and languages had a higher concentration of users, providing insights for targeted marketing.