**Comprehensive Documentation: Data Generation Process Using Mockaroo**

**Objective**: Generate realistic user data (60,000+ records) for testing, mimicking production data structures.

**1. Tool & Setup**

**a. Mockaroo Overview**

* **Website**: **[mockaroo.com](https://www.mockaroo.com/" \t "_blank)**
* **Key Features**:
  + Generate up to **1,000 rows** for free; **$50/year** for expanded limits.
  + Output formats: CSV, JSON, SQL, Excel.
  + Docker deployment for private cloud integration.
  + Intuitive UI with drag-and-drop schema design.

**b. Workflow Summary**

1. **Schema Design**: Define fields, data types, and validation rules.
2. **Data Generation**: Configure row count and output format.
3. **Export**: Download data for testing or integration.

A screenshot of a computer

AI-generated content may be incorrect.

**2. Step-by-Step Process**

**Step 1: Schema Definition**

* **Fields & Data Types** (Based on Data Preview.png):

| **Field Name** | **Data Type** | **Configuration** |
| --- | --- | --- |
| Id | Row Number | Sequential unique identifier. |
| first\_name | First Name | Gender-specific names enabled. |
| last\_name | Last Name | Randomized surnames. |
| CountryName | Country | Specific countries selected (e.g., Brazil, China). |
| State | State (US/CA) | Related to CountryName (e.g., Illinois for USA). |
| City | City | Randomized based on CountryName and State. |
| App | Number (Range) | Age simulation (min: 20, max: 60). |
| Gender | Gender (Abbrev) | Options: M, F, W (custom "W" for testing). |
| Email | Email Address | Custom domains (e.g., flygah.com, adobe.com). |
| Password(msg) | Password Hash | Bcrypt format with salt ($2a$...). |

* **Advanced Settings**:
  + **Blank Values**: Configured 10% blanks for State and City to simulate missing data.
  + **Formats**:
    - **Password**: $2a$10$[salt][hash] (bcrypt algorithm).
    - **Email**: Custom domains for realism (e.g., @flygah.com).

**Step 2: Data Generation**

* **Volume**: Generated **60,000 rows** (paid plan required for >1,000 rows).
* **Preview**: Used **PREVIEW** button to validate data structure.
* **Output Format**:
  + **CSV** with Unix line endings (LF).
  + Included header row for clarity.

**Step 3: Export & Post-Processing**

* **File Name**: Final Result of data Generation.csv.
* **Data Cleaning** (Post-Export):
  + **Gender Consistency**: Fixed mismatches (e.g., Gender column had Male/Female alongside M/F).
  + **Email Validation**: Removed invalid emails (e.g., dwell19bombo.com missing @).
  + **Country Fixes**: Corrected typos (e.g., JADAA → JAPAN).

**3. Final Dataset Structure**

**Key Columns** (Final Result of data Generation.png):

| **Column** | **Type** | **Example Value** | **Purpose** |
| --- | --- | --- | --- |
| Id | Integer | 1 | Unique user identifier. |
| Exercise - Last Name | Text | Bandy | Simulated user activity. |
| Stated - All - Gender | Text | 25 F | Age + gender combo. |
| E-mail | Text | adandry4@theatlantic.com | Test login credentials. |
| Cessation | Float | 18520979174 | Simulated transaction ID. |
| Notifications | Datetime | 2023-09-29 00:00:00 | User notification timestamps. |
| Limited | Datetime | 2025-04-04 00:01:36 | Account expiration date. |

**4. Challenges & Solutions**

**a. Data Consistency**

* **Issue**: Mismatched gender abbreviations (F vs. Female).
* **Solution**: Used Power Query to standardize values using Text.Replace.

**b. Invalid Emails**

* **Issue**: Missing @ in emails (e.g., dwell19bombo.com).
* **Solution**: Added validation rule in Mockaroo:

regex

Copy

Download

^[A-Za-z0-9.\_%+-]+@[A-Za-z0-9.-]+\.[A-Za-z]{2,}$

**c. Country Typos**

* **Issue**: Fictional countries (e.g., JADAA, TENALAND).
* **Solution**: Restricted CountryName to a predefined list in Mockaroo.

**d. Large Volume Handling**

* **Issue**: Generating 60,000 rows caused browser lag.
* **Solution**: Used **Docker image** for local deployment and batch processing.

**5. Mockaroo Interface Breakdown**

**a. Top Banner Area**

* **Key Messages**:
  + "Generate up to 1,000 rows of realistic test data."
  + Paid plans start at **$50/year** for higher limits.
  + Docker integration for private cloud usage.

**b. Main Data Definition Table**

* **Dynamic Controls**:
  + **Blank %**: Set null rates for fields (e.g., 10% for State).
  + **Preview** (🔍): Sampled data for validation.
  + **Advanced Settings** (⚙️): Custom regex for emails/phones.

**c. Output Configuration**

* **Rows**: 60,000 (user-defined).
* **Format**: CSV with headers.
* **Buttons**:
  + **GENERATE DATA** (green): Initiated data creation.
  + **SAVE AS**: Stored schema for reuse.

**6. Ethical & Practical Considerations**

* **Data Privacy**: Avoided real user information; all data synthetic.
* **Formatting**: Ensured GDPR-compliant pseudonymization (e.g., hashed passwords).
* **Performance**: Optimized CSV exports for database integration.

**7. Results & Applications**

* **Volume**: 60,000 rows across 28 fields.
* **Use Cases**:
  + Load testing for user authentication systems.
  + Database stress testing.
  + UI/UX testing for e-learning platforms.
* **Anomalies**:
  + Cessation column included scientific notation (e.g., 8.6756E1-12), fixed via Excel formatting.

A screenshot of a computer

AI-generated content may be incorrect.