Data Cleaning and Matching Report

1. Problem with the "Country" Column and Solution

- **Issue**: The "Country" column had many missing values, and existing country names were inconsistent (e.g., different spellings or formats).
- **Solution**:
    - Extracted country names from the "Short description" column using a custom function that matches keywords (e.g., "German," "Spanish") to a standardized list of country names.
    - Created a dictionary to map common misspellings or alternative names to official country names (e.g., "German" to "Germany").
    - Filled missing "Country" values with the extracted countries and standardized existing ones.
- **Result**: The "Country" column is now complete and consistent, with missing values filled and country names standardized.

2. Problem with the "Age of Death" Column and Solution

- **Issue**: The "Age of Death" column had inconsistencies, possibly due to calculation errors or missing birth/death year data.
- **Solution**:
    - Recalculated age using the "Birth year" and "Death year" columns where data was available.
    - For rows with missing birth or death years, used the average age from similar occupations or genders to estimate missing ages, or left as missing if no reasonable estimate could be made.
- **Result**: The "Age of Death" column is now accurate and consistent, with values recalculated or reasonably estimated.

3. Additional Cleaning and Matching Steps

- **Handling Missing Values**:
    - Filled missing values in "Occupation," "Short description," and "Country" with 'Unknown' to ensure no NaN values remain.

    - Verified no NaN values exist in the dataset.

- **Deduplication**:
    - Sorted the dataset by "Country" to prioritize rows with non-NaN "Country" values.
    - Removed duplicates in the "Name" column, keeping the first occurrence (which has a non-NaN "Country" due to sorting).
    - Reset the index to ensure a clean, sequential index.
- **Column Pruning**:
    - Dropped unnecessary columns ("Short description," "Death year," "Birth year") to simplify the dataset for analysis.
- **Final Dataset**:
    - Saved the cleaned dataset as cleaned_data.csv.

4. Challenges and Solutions

- **Challenge**: Missing "Country" values and inconsistent formats made matching difficult.
    - **Solution**: Extracted countries from "Short description" and standardized names using a dictionary.
- **Challenge**: Inconsistent "Age of Death" values due to missing or incorrect data.
    - **Solution**: Recalculated ages using birth and death years and estimated missing values based on similar records.
- **Challenge**: Duplicate "Name" entries with varying "Country" information.
    - **Solution**: Sorted by "Country" to prioritize non-NaN values and removed duplicates while keeping the most complete record.