Assessment Task: Data Deduplication, Matching, and Analysis

Overview

In this assessment, you are provided with two datasets, **Dataset 1** and **Dataset 2**, Your task is to:

1. Perform **exploratory data analysis (EDA)** on both datasets to understand their structure, inconsistencies, and distributions.
2. Preprocess the datasets to clean, standardize, and prepare them for matching.
3. Use data matching techniques to accurately identify and link related records between the two datasets.
4. Conduct data analysis to answer specific questions that can only be addressed after the matching process is complete.

Datasets

- **Dataset 1:** Contains the following fields:
  - `Name`
  - `Short description`
  - `Gender`
  - `Occupation`
  - `Age of death`

- **Dataset 2:** Contains the following fields:
  - `Full_name`
  - `Gender`
  - `Country`
  - `Birth year`
  - `Death year`
  - `Manner of death`

Task Instructions

Part 1: Exploratory Data Analysis (EDA)

Perform an initial analysis of both datasets to:

- Understand the distribution of key fields such as `Gender`, `Country`, and `Occupation`.

- Identify missing values and inconsistencies.

Part 2: Data Preprocessing

- Clean and standardize fields like `Name`, `Country`, and `Gender` to align them across both datasets.
- Handle missing values and remove irrelevant records if necessary.
- Generate new columns or features to aid in the matching process (e.g., similarity scores or tokenized fields).

Part 3: Data Matching

- Use appropriate data matching techniques (e.g., exact matching, fuzzy matching, or rule-based matching) to deduplicate records within each dataset and match records across the two datasets.
- Ensure a high level of accuracy by addressing common typos and logical inconsistencies.
- Provide a mapping of matched records, including the IDs or key fields that link the records.

Part 4: Analysis and Insights

Once the matching is complete, answer the following questions:

1. **Gender Distribution:**
   - What is the distribution of genders (`Male` and `Female`) in the combined dataset? How does it differ between the two datasets before matching?
2. **Geographic Representation:**
   - Which countries are most represented in the dataset, and are there any discrepancies in country distributions between Dataset 1 and Dataset 2?
3. **Occupation Trends:**
   - What are the most common occupations, and how do they vary by gender?
4. **Historical Patterns:**
   - What is the average age at death for individuals in the dataset? Are there differences based on gender or country?
5. **Analysis of Missing Data:**
   - After matching, what proportion of records still contain missing values in key fields such as `Country` or `Manner of death`? What strategies could be used to impute or address these gaps?

Part 5: Dashboard Development

Using Power BI, create a comprehensive dashboard based on the matched data to provide visual insights for stakeholders.

Deliverables

- **Cleaned Datasets:** Provide the cleaned and matched versions of Dataset 1 and Dataset 2.
- **Matching Report:** Include a summary of the matching process, highlighting challenges and solutions.
- **Analysis Report:** Provide answers to the analysis questions with supporting visualizations or tables.
- **Provide the Power BI file** (.pbix) along with a user guide to navigate the dashboard. Include screenshots of key visualizations in the report to ensure stakeholders can quickly understand the insights.
- **Code and Methodology:** Submit the scripts or notebooks used for EDA, preprocessing, matching, and analysis, with comments explaining key steps.

Evaluation Criteria

Your submission will be evaluated based on:

- Accuracy and efficiency of the matching process.
- Clarity and completeness of the analysis.
- Quality of data cleaning and preprocessing.
- Insightfulness of answers to the analysis questions.
- Code readability and documentation.

Good luck, and happy analyzing!