Data Report

Question

Does having a higher academic degree correlate with higher income in the Americas?

Data Sources

Dataset 1: American Citizen Income Dataset

Source: Kaggle: American Citizen Income

Metadata URL: https://www.kaggle.com/datasets/amirhosseinmirzaie/americancitizenincome

Description: Provides individual-level data on age, education, native country, and income classification (<=50K, >50K) for residents of the Americas.

Structure: Tabular format with categorical and numerical columns.

Quality: Contains missing and ambiguous values (e.g., ?), requiring cleaning.

License: Provided under Kaggle's terms of use for educational purposes.

Dataset 2: Global Salary DataSet 2022

Source: Kaggle: Global Salary DataSet 2022

Metadata URL: https://www.kaggle.com/datasets/ricardoaugas/salary-transparency-dataset-2022

Description: Aggregates country-level salary data by education levels, age ranges, and salary brackets.

Structure: Tabular format with columns for age ranges, education levels, countries, and salary categories.

Quality: No missing values, consistent formatting, but requires alignment with Dataset 1's structure for effective integration.

License: Provided under the Community Data License Agreement (CDLA). For details, visit: https://cdla.dev/sharing-1-0/.

Data Pipeline

The pipeline automates data extraction, cleaning, integration, and storage. It processes two datasets to ensure compatibility and creates a unified dataset for analysis.

Technologies Used:

- Programming Language: Python
- Libraries: Pandas, SQLite
- Pipeline Script: pipeline.py
- Execution: pipeline.sh (Bash script to run the pipeline)

Steps Performed:

1. Data Extraction:
   - Dataset 1 was downloaded from Kaggle using the Kaggle API and loaded as a CSV file.
   - Dataset 2 was manually downloaded due to limited API access and loaded into the pipeline.
2. Data Cleaning:
   - Removed rows with missing values (?) and ensured consistent formatting.
   - Dropped irrelevant columns (fnlwgt, race, sex, etc.) from Dataset 1.
3. Data Transformation and Integration:
   - Aligned column names and merged the datasets on the native.country/Country fields.
   - Transformed numerical salary data from Dataset 2 into categorical income levels (<=50K or >50K) for consistency.
4. Data Storage:
   - Stored the unified dataset in the /data directory as an SQLite file (data.sqlite).

Challenges and Solutions

1. Aligning different column formats between the datasets:
   - Solution: Standardized column names and used Pandas to merge the datasets effectively.
2. Handling missing and ambiguous values in Dataset 1:

- Solution: Dropped rows with ? and ensured consistent formatting.

Results and Limitations

Output Data:

- Structure: Combined dataset with columns: age, education, country, income.
- Quality: Cleaned, formatted, and free of missing values. Normalized and aligned to ensure comparability across datasets.

Format:

- Stored as an SQLite database (data.sqlite) for easy querying and scalability.

Limitations:

- Dataset 2's country-level aggregation introduces potential inconsistencies when integrated with individual-level data from Dataset 1.
- Some countries in Dataset 1 may lack representation in Dataset 2, leading to partial gaps in analysis.
- Income data in Dataset 2 is estimated and may not reflect actual earnings.