Data Quality Analysis Report: Adult Salary Dataset (Himanshu, MDS202327)

Integrated Findings from YData Profiling, PyDeequ, and Great Expectations

1. Overview

This report consolidates data quality findings from three tools—YData Profiling, PyDeequ, and Great Expectations—applied to the Adult Salary dataset (32,560 records, 15 variables). The tools were used to cross-validate and deepen the understanding of data quality dimensions, including completeness, consistency, uniqueness, and distributional properties.

2. Findings by Tool

A. YData Profiling

- **Completeness:** No missing values in any column; 0% missing cells.
- **Duplicates:** 23 duplicate rows (0.1%).
- **Distributional Issues:**
  - `capital_gain` (91.7% zeros) and `capital_loss` (95.3% zeros) are highly sparse.
  - `race` (85.4% White) and `native_country` (89.6% United States) are highly imbalanced.
- **Correlations:**
  - `education` and `education_num` are perfectly correlated.
  - `relationship` and `sex` are also highly correlated.
- **Special Values:**
  - "?" used in `workclass`, `occupation`, and `native_country` to denote unknowns.
- **Cardinality:**
  - `fnlwgt` has 21,647 distinct values (66.5% of records), indicating it is not a unique identifier.
- **No infinite or negative values** in numeric columns.
- **Class Imbalance:**
  - `income` is skewed (75.9% <=50K, 24.1% >50K)

B. PyDeequ

- **Constraint Validation:**
  - Confirmed all columns are non-null.
  - Detected that categorical columns (`workclass`, `occupation`, `native_country`) contain the "?" value, flagging these as pseudo-missing.
  - Asserted expected value ranges for numeric columns (e.g., `age` between 17 and 90, `hours_per_week` between 1 and 99), all passing.
- **Uniqueness:**
  - Verified that `fnlwgt` is not unique.
- **Distribution Checks:**
  - Detected high sparsity in `capital_gain` and `capital_loss`.
- **Duplicate Detection:**
  - Independently confirmed presence of duplicate rows.
- **Category Cardinality:**
  - Detected high cardinality in `native_country` and `occupation`, but with dominant categories.

C. Great Expectations

- **Expectation Suites:**
  - Validated that all columns have non-null values.
  - Explicitly checked for and flagged the "?" value as a failing expectation for true completeness.
  - Validated value ranges for numeric columns, all passing.
- **Distributional Expectations:**
  - Created expectations for class balance in `income`, `race`, and `native_country`; flagged significant imbalance.
- **Duplicate Row Expectation:**
  - Confirmed presence of duplicates, matching other tools.
- **Correlations and Redundancy:**
  - Noted perfect correlation between `education` and `education_num`; recommended dropping one.

3. Comparison of Tools and Cross-Validation

Aspect	YData Profiling	PyDeequ	Great Expectations
Missing values	0% missing	All columns non-null	All columns non-null
Special values ("?")	Detected, descriptive only	Flagged as pseudo-missing	Explicitly failed expectation
Duplicates	23 rows (0.1%)	Confirmed	Confirmed
Numeric ranges	Descriptive stats	Validated constraints	Validated expectations
Cardinality	Described for all columns	Checked for uniqueness/high	Checked for valid categories
Distribution imbalance	Highlighted	Detected	Explicitly flagged
Correlations	Quantified, flagged	Noted in summary	Noted, recommended action

Remarks on Comparison:

- All three tools agreed on the absence of true missing values, presence of duplicates, and the dominance of certain categories in categorical columns.
- PyDeequ and Great Expectations were able to explicitly flag the "?" value as a data quality concern, while YData Profiling only described its presence.
- Only YData Profiling provided a detailed correlation matrix, while the other tools noted redundancy as part of rule-based checks.
- All tools validated numeric ranges and flagged no out-of-bounds values.
- Distributional imbalance was highlighted by all, but Great Expectations allowed for explicit expectation failures on class balance.

4. Summary Table: Key Data Quality Issues

Issue	YData Profiling	PyDeequ	Great Expectations	Remarks
Missing values	None	None	None	True completeness
Special value ("?")	Described	Flagged	Explicitly failed	Needs imputation
Duplicate rows	23	23	23	Remove for modeling
Class imbalance	High	High	High	Consider resampling
Numeric outliers/sparsity	High zeros	High	High	Affects modeling
Redundant features	Noted	Noted	Action recommended	Drop one of pair

5. Conclusions and Recommendations

- **All tools agree** the dataset is generally clean, with no missing values or out-of-range numerics.
- **Duplicates and special values ("?")** are a consistent data quality concern and should be addressed before analysis.
- **Class imbalance** and **high sparsity** in some numeric columns may affect downstream tasks.
- **Redundant features** (e.g., `education` and `education_num`) should be reviewed for potential removal.
- **Cross-tool validation** strengthens confidence in findings and highlights the value of combining descriptive and rule-based data quality checks.