📘 Data Quality Metrics Report

Name: SHREYA SINGH
Roll Number: MDS202434
Dataset: UCI Adult Dataset
Tools Used: PyDeequ, YData Profiling, Great Expectations

📊 Data Quality Metrics Summary

✅ 1. Metrics from PyDeequ

Column	Metric	Value
workclass	Completeness	1.0
education	ApproxCountDistinct	16
hours_per_week	Mean	40.44
AgeAbove18	Compliance	0.988
age	Completeness	1.0

Highlights:

- Full completeness for `workclass` and `age`.
- `education` has 16 distinct values, indicating diverse education levels.
- `hours_per_week` average aligns with full-time work expectations.
- High compliance with adult age condition (`AgeAbove18 > 0.98`).

✅ 2. Findings from Great Expectations

Expectation Suite Overview:

- Expectations were created using both profiler and manual additions.
- The suite includes type validation, null checks, value ranges, and distribution checks.

Notable Validations:

- ✅ `expect_column_values_to_be_between` for `age`: all values fall between 17 and 90.
- ✅ `expect_column_values_to_not_be_null` passed for `age`, `hours-per-week`, etc.
- ✅ `expect_column_median_to_be_between`: median of `age` and `hours-per-week` within realistic bounds.
- ✅ `expect_column_values_to_match_regex` applied to string fields like `occupation`, confirming expected naming format.

Warnings / Issues Detected:

- ⚠️ `native-country`, `workclass`, and `occupation` showed potential nulls or unexpected distinct values.
- ⚠️ Slight variation from expected categorical distributions in `relationship`, `education`, and `race`.
- ⚠️ Columns with skewed numerical values (e.g., `capital-gain`) had outliers not caught by basic validation.

Useful Metadata:

- Great Expectations tracks batch ID, expectation success rates, and the number of evaluated rows.
- Validation run produced a full data documentation bundle (checkpoints, suites, validations).

✅ 3. Insights from YData Profiling

The YData Profiling report performed deep statistical and visual profiling of all columns.

Top-Level Summary:

- Variables: 15
- Total Rows: ~32,561
- Missing Cells: ~3.4% (mainly in `workclass`, `occupation`, `native-country`)

Categorical Insights:

- `education` had 16 unique values; top 3: HS-grad, Some-college, Bachelors.
- `marital-status`, `relationship`, and `occupation` showed moderate cardinality and expected distributions.

Numerical Stats:

- `age`: Mean ~38.6, Std ~13.6, min 17, max 90 (normal distribution)
- `hours-per-week`: Right-skewed, with a sharp spike at 40 hours.
- `capital-gain` and `capital-loss`: Strong right skew with many zero values (outliers identified visually).

Correlations (Pearson & Spearman):

- High positive correlation between `education-num` and `education`.
- Mild negative correlation between `age` and `hours-per-week`.
- Potential multicollinearity risk among derived features.

Warnings & Flags:

- ⚠️ Highly imbalanced target variable (`income` class).
- ⚠️ Strong skewness in financial columns could affect modeling accuracy.
- ⚠️ Zero-inflated distributions in `capital-gain/loss`.

📌 Combined Takeaways & Data Quality Action Plan

Issue Area	Description	Suggested Action
Missing Data	`workclass`, `occupation`, `native-country` have nulls	Impute or mark missing values
Skewed Distributions	`capital-gain`, `capital-loss`, `hours-per-week` are heavily skewed	Apply normalization or transformation
Outliers	Found in weekly hours and income-related columns	Winsorization or outlier handling methods
Schema Drift	Minor categorical value inconsistencies detected	Review categorical encoding rules
Compliance	High compliance with age and data type expectations	Maintain data validation logic

---

📁 Attached References

- ✅ PyDeequ Metrics Table (CSV)
- ✅ `great_expectations.pdf` — GE Notebook Output
- ✅ `ydata_profile.pdf` — Full EDA Report