Data Quality Report

1.Dataset Source & Intent

The dataset contains **205 student-level records with 16 features**, comprising academic and demographic information collected for analysis and modeling purposes. The primary intent of the dataset is to support exploratory data analysis and the development of predictive models related to student performance and outcomes. Prior to modeling, the data required systematic cleaning to address quality issues, inconsistencies, and violations of defined business rules.

2. Data Quality Issues

Issue 1: Placeholder Missing Values

- **Problem:** Missing data was represented using placeholders such as "NA", "N/A", "null", "unknown", and empty strings.
- **Action Taken:** All placeholder values were standardized and converted to proper missing values (NaN).
- **Code Reference:** Missing indicator standardization step.
- **Rationale:** Ensures consistent missing value handling across columns.
- **Effect:** Reduced inconsistent representations of missing data; improved reliability of imputation.

Issue 2: Invalid and Mixed Date Formats

- **Problem:** Date columns contained invalid formats and non-date strings.
- **Action Taken:** Dates were parsed using pd.to_datetime(errors="coerce"), converting invalid entries to NaT.
- **Code Reference:** Date parsing step.
- **Rationale:** Prevents parsing errors and enables time-based logic.
- **Effect:** Invalid dates converted to NaT; count of invalid dates was logged.

Issue 3: Inconsistent Categorical Values

- **Problem:** Categorical fields (gender, course stream, device type, scholarship, internet access) had inconsistent casing and multiple representations.
- **Action Taken:** Categories were standardized using custom mapping functions and normalized casing.
- **Code Reference:** Categorical standardization step.
- **Rationale:** Improves interpretability and ensures consistent encoding.
- **Effect:** Reduced category cardinality and eliminated ambiguous labels.

Issue 4: Duplicate Records

- **Problem:** The dataset contained both exact duplicate rows and multiple records per student.
- **Action Taken:**
  - Exact duplicates were removed.
  - For student-level duplicates, the most recent record (based on date) was retained.
- **Code Reference:** Duplicate handling step.
- **Rationale:** Ensures one valid record per student.
- **Effect:** Total duplicate rows reduced; final dataset contains unique student entries.

Issue 5: Business Rule Violations

- **Problem:**
  - Students older than 17 were assigned non-degree programs.
  - Students younger than 17 were assigned degree-level programs.
- **Action Taken:**
  - Enforced age-based course stream rules.
  - Invalid course stream values were set to missing (NaN) for later imputation.
- **Code Reference:** Business rules enforcement step.
- **Rationale:** Aligns data with realistic academic constraints.
- **Effect:** All remaining age–course stream combinations are valid after imputation.

3. Assumptions Made

- The most recent student record best represents the current student status.
- Median imputation is appropriate for numeric variables to reduce outlier influence.
- Mode imputation is appropriate for categorical variables.
- Degree programs are valid only for students older than 17.
- Invalid values should be corrected rather than removing records to preserve dataset size.

4. Final Missingness Summary

- After cleaning and business rule enforcement, missing values were imputed:
  - **Numeric columns:** Median
  - **Categorical columns:** Mode
- A final missingness table (top 10 columns by missing count) was reviewed.
- No critical modeling columns were left with missing values.
- Some categorical values may reflect imputed "Unknown"-type categories by design.

5. Final Outcome

The final cleaned dataset is consistent, rule-compliant, and suitable for analysis. A separate encoded, model-ready dataset was produced for machine learning tasks. All cleaning steps are reproducible and documented in the preprocessing notebook.