

# ■ Data Quality Report

## 1. Context

The student dataset originally contained around 205 rows and 16 columns. The objective was to clean and prepare the data so that it could be reliably used for exploratory analysis (EDA) and modeling tasks. The process focused on identifying and fixing common data quality issues.

## 2. Top Issues Found and Fixed

### 1. Mixed Date Formats (`admission_date`)

Dates appeared in multiple formats such as 11-08-2023, 01-15-2024, and Aug 19, 2025. All dates were standardized to the format dd-mm-yyyy.

### 2. Currency Strings in `fee_paid_inr`

The fee column contained ■ symbols, commas, and text. All non-numeric characters were removed and values were converted to numeric type. Invalid entries were turned into NaN.

### 3. Placeholder Text for Missing Values

Values like 'NA', 'N/A', 'Unknown', and '-' were replaced with proper NaN. This allowed for clear identification of missing data, especially in scholarship, parental\_education, and course\_stream.

### 4. Categorical Inconsistencies

Columns such as gender, has\_internet, and device\_type contained inconsistent entries (e.g., 'M', 'male', 'MALE'). These were standardized using mapping dictionaries, reducing messy variations to clean categories.

### 5. Duplicate Student Records

Some students had multiple rows. Exact duplicates were removed, and for repeated `student_id` values, the row with the most recent `admission_date` was retained. The dataset was reduced from 205 rows to about 200 unique student records.

## 3. Post-Cleaning Status

- Main fields now contain no missing values.
- The top 10 columns show 0% missingness.
- Parental\_Education and Scholarship were left as "Unknown" where data was missing. In reporting, these are treated as missing values, but in the dataset they remain explicit categories so that no rows are dropped.