Data Quality Report

Context:

I worked on cleaning the student dataset (STUDENTS_RAW_DATA.csv) in my notebook (*Lohith_Task1_Task2.ipynb*).
It originally had around 205 rows × 16 columns of student information.
The main aim was to prepare this raw data so that it can be properly used for exploratory analysis (EDA) and modeling tasks.

Top 5 Issues I Found and Fixed

1. **Mixed date formats (admission_date)**
    - The admission dates were written in different ways like 11-08-2023, 01-15-2024, Aug 19, 2025.
    - I used pd.to_datetime with both dayfirst=True/False and fallback parsing to standardize them.
    - Now all valid dates are stored in the format dd-mm-yyyy.
2. **Currency strings in fee_paid_inr**
    - The fee column had ₹ symbols, commas, and text mixed in.
    - I removed everything except numbers using regex and converted it to numeric type.
    - Some invalid values turned into NaN, but it's now a proper numeric column.
3. **Placeholder text for missing values**
    - Columns had things like "NA", "N/A", "Unknown", or "-" instead of blanks.
    - I replaced all such placeholders with proper NaN.
    - This helped me clearly see where data was missing (especially in scholarship, parental_education, and course_stream).
4. **Categorical inconsistencies**
    - Columns like gender, has_internet, and device_type had inconsistent entries (e.g., "M", "male", "MALE", "Laptop" vs "laptop").
    - I standardized these using mapping dictionaries so that only clean values remain.
    - For example, gender was reduced from 7 messy variations down to 4 clean categories.
5. **Duplicate student records**
    - Some students had multiple rows.
    - I first removed exact duplicates, and for cases with the same student_id I kept the row with the latest valid admission_date.
    - This brought the dataset from 205 rows down to about 200 unique student records.

After cleaning, the dataset no longer has missing values in the main fields.
The top 10 columns all show **0% missingness**(can refer the notebook file)

Note:

- Parental_Education and Scholarship were intentionally kept as "Unknown" where data was missing.
- For reporting, "Unknown" is treated as missing, but in the cleaned dataset these are left as explicit categories so that no rows are dropped.