# Data Quality Report

### Dataset Overview:

The dataset is sourced from the UCI Student Performance Dataset and contains 649 rows with 33 features.  
The target variable is G3 (final grade), a numeric feature ranging from 0 to 20.  
Features include demographics, study habits, school support, attendance, and prior grades (G1 and G2).

### Schema and Types:

Below is a table summarizing the types and description of each feature

This table was generated using python.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **name** | **role** | **type** | **demographic** | **description** | **units** | **missing values** |
| **0** | school | Feature | Categorical | None | student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira) | None | no |
| **1** | sex | Feature | Binary | Sex | student's sex (binary: 'F' - female or 'M' - male) | None | no |
| **2** | age | Feature | Integer | Age | student's age (numeric: from 15 to 22) | None | no |
| **3** | address | Feature | Categorical | None | student's home address type (binary: 'U' - urban or 'R' - rural) | None | no |
| **4** | famsize | Feature | Categorical | Other | family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3) | None | no |
| **5** | Pstatus | Feature | Categorical | Other | parent's cohabitation status (binary: 'T' - living together or 'A' - apart) | None | no |
| **6** | Medu | Feature | Integer | Education Level | mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education) | None | no |
| **7** | Fedu | Feature | Integer | Education Level | father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education) | None | no |
| **8** | Mjob | Feature | Categorical | Occupation | mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other') | None | no |
| **9** | Fjob | Feature | Categorical | Occupation | father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other') | None | no |
| **10** | reason | Feature | Categorical | None | reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other') | None | no |
| **11** | guardian | Feature | Categorical | None | student's guardian (nominal: 'mother', 'father' or 'other') | None | no |
| **12** | traveltime | Feature | Integer | None | home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour) | None | no |
| **13** | studytime | Feature | Integer | None | weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) | None | no |
| **14** | failures | Feature | Integer | None | number of past class failures (numeric: n if 1<=n<3, else 4) | None | no |
| **15** | schoolsup | Feature | Binary | None | extra educational support (binary: yes or no) | None | no |
| **16** | famsup | Feature | Binary | None | family educational support (binary: yes or no) | None | no |
| **17** | paid | Feature | Binary | None | extra paid classes within the course subject (Math or Portuguese) (binary: yes or no) | None | no |
| **18** | activities | Feature | Binary | None | extra-curricular activities (binary: yes or no) | None | no |
| **19** | nursery | Feature | Binary | None | attended nursery school (binary: yes or no) | None | no |
| **20** | higher | Feature | Binary | None | wants to take higher education (binary: yes or no) | None | no |
| **21** | internet | Feature | Binary | None | Internet access at home (binary: yes or no) | None | no |
| **22** | romantic | Feature | Binary | None | with a romantic relationship (binary: yes or no) | None | no |
| **23** | famrel | Feature | Integer | None | quality of family relationships (numeric: from 1 - very bad to 5 - excellent) | None | no |
| **24** | freetime | Feature | Integer | None | free time after school (numeric: from 1 - very low to 5 - very high) | None | no |
| **25** | goout | Feature | Integer | None | going out with friends (numeric: from 1 - very low to 5 - very high) | None | no |
| **26** | Dalc | Feature | Integer | None | workday alcohol consumption (numeric: from 1 - very low to 5 - very high) | None | no |
| **27** | Walc | Feature | Integer | None | weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) | None | no |
| **28** | health | Feature | Integer | None | current health status (numeric: from 1 - very bad to 5 - very good) | None | no |
| **29** | absences | Feature | Integer | None | number of school absences (numeric: from 0 to 93) | None | no |
| **30** | G1 | Target | Categorical | None | first period grade (numeric: from 0 to 20) | None | no |
| **31** | G2 | Target | Categorical | None | second period grade (numeric: from 0 to 20) | None | no |
| **32** | G3 | Target | Integer | None | final grade (numeric: from 0 to 20, output target) | None | no |

### Duplicates:

No duplicates were found.

### Outlier Detection and handling:

Using the IQR method for outlier detection and handling I came to the following conclusions

Looking at the features, absences at first glance look a little off but after some thinking , yes some students can have been absent for most of the year (Reasons could be Travel) but to not let it affect the model, we can replace the values beyond a threshold with the nearest non-outlier value.

A graph with a box plot

AI-generated content may be incorrect.

Now for features like Famrel , FreeTime and Dalc (Daily alcoholic consumption) have some outliers, but they are not significant and are also logical.

A blue rectangle and white rectangular shapes

AI-generated content may be incorrect.

Finally, the failures feature, at first glance, looks like it needs to be dropped right away but it is an ordinal Categorical value with values ranging from 0 (no past failures) to 3 (3 or more past failures).

A graph with a number of small circles

AI-generated content may be incorrect.

I capped absence to the 95th percentile for the top 5% values which are greater than 30

Now let’s look at the before and after of the absence feature.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Min Value | Max Value | Lower Bound | Upper Bound | Outlier Count |
| Before capping | 0 | 32 | -9.0 | 15.0 | 21 |
| After Capping | 0 | 12 | -9.0 | 15.0 | 0 |

Note that the lower bound being negative is just the formula, but if you look at the boxplot you will see that there are no negative values.

### Final Summary:

Overall, the dataset is clean, with no duplicates or missing values. The main data quality issue was outliers in the absences feature, identified using the IQR method and handled via capping. Another consideration for modeling is the potential leakage from G1 and G2 (previous grades) on the target G3. This will be addressed by training models both with and without these features.