These are the notes on the data quality of the train.csv and the test.csv files of the Mental Health Issue Prediction competition.

Completeness:

|  |  |
| --- | --- |
| Column | Number of empty entries |
| Profession | 36630 |
| Academic Pressure, Study Satisfaction | 112803 |
| Work Pressure | 27918 |
| CGPA | 112802 |
| Job Satisfaction | 27910 |
| Degree | 2 |
| Dietary Habits, Financial Stress | 4 |

* Majority of entries in Academic Pressure, Study Satisfaction and CGPA are empty and rows with empty values in Academic Pressure tend to be empty in Study Satisfaction and CGPA
* Entries that are empty in Work Pressure are also tend to be empty in Job Satisfaction and Profession

Consistency:

* Some entries in CGPA were formatted as decimals and some like whole numbers
* Name, City, Dietary Habits and Degree column all had entries that were numbers even though the majority of entries in all of those columns were text

Accuracy:

* Name column has six entries with wrong values like 18, M. Tech (which is a type of degree) and at times the incorrect value is the same as the profession or degree type
* City column has 11 entries with wrong values like 3, City, less than 5 Kalyan (will be changed to Kalyan) and degree types
* Profession column has 22 entries with wrong values like degree types, names of subjects and city names
* Sleep duration has only one entry with a wrong value that is a date
* Dietary habits column contains at least 17 entries with wrong values like numbers, degree types, yes and no values and possibly duplicate values in not healthy and unhealthy and less healthy and less than healthy
* Degree column has multiple entries with wrong values like numbers, certain professions, names, B. Student and so on

Uniqueness:

* There are no duplicate rows