**Assessment 1 – report**

Task 1 - Examination of the data types assigned by for each variable by the pandas library.

* The pandas library was imported, and the data read into memory with the pandas read\_csv() method.
* The dataset was examined using the pandas info() method to obtain information about the variable datatypes.
* Each variable was examined in turn using the pandas value\_counts() method. The provided a series object showing the frequency of each distinct row in the data set. The results of these method calls are shown in Figure 1.
* Mismatches between expected datatypes read by pandas and those described in the dataset description are shown in Table 1, with the approach to correct the data types.

A screenshot of a computer

Description automatically generated

Figure 1 - results from pandas info(), and value\_counts() methods on the loaded D1 dataset.

The output from the method calls shows that some variables are loaded as string types (pandas objects) when the data description identifies these as numeric types e.g. number\_outpatient, number\_inpatient, number\_emergency. These examples are loaded as string due to the presence of a ‘?’ character to denote missing/unknown values.

Table 1 - Table of mismatched data types and proposed corrections

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Variable Name** | **Description** | **Pandas data type** | **Mismatch corrective action** |
| 16 | number\_outpatient | Number of outpatient visits of the patient in the year preceding the encounter | Object type (string) | There are 20 missing values represented by ‘?’.  Convert the ? chars to NaNs and convert the column to int64 |
| 17 | number\_emergency | Number of emergency visits of the patient in the year preceding the encounter | Object type (string) | There are 69 missing values represented by ‘?’.  Convert the ? chars to NaNs and convert the column to int64 |
| 18 | number\_inpatient | Number of inpatient visits of the patient in the year preceding the encounter | Object type (string) | There are 15 missing values represented by ‘?’.  Convert the ? chars to NaNs and convert the column to int64 |

**Comments about the goal of the data mining**

* This looks like a 'length of stay' prediction problem. The goal is to predict the length of stay of a patient in the hospital. The `length\_of\_stay` column is the target variable. It has no missing values and the data are in a manageable range. We should convert this column to a numeric data type.
* The `readmitted` column could be secondary target variable. It is a categorical variable with three classes. We should convert this column to a categorical data type.
* The `discharge\_disposition\_id` could also be used as a secondary target variable. It is a categorical variable with 26 classes. It might be worth reducing the number of classes to binary outcome variable (all cause mortality), or categorical variable with fewer classes (e.g. discharged home, discharged to another facility, died.).
* We should discuss if we want to filter out the `admission\_type\_id` column. If we choose length of stay as the target variable, we might want to filter out the `admission\_type\_id` column to exclude newborns and electives. The same goes for `single\_day\_admission`. We might want to filter out the single day admissions.