**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 three variables were loaded as string types (pandas object type), whereas the data description identified these as numeric types The variables are: number\_outpatient, number\_inpatient, number\_emergency. These were 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** |
| 5 | Age | Age quantile | Object type (string) | The data were intended to be in deciles and should be treated as an ordinal variable.  Convert to an ordered qualitative data type. |
| 6 | Weight | Weight in ponts | Object type (string) | The data were intended to be in deciles and should be treated as an ordinal variable.  Convert to an ordered qualitative data type. |
| 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 |

The mismatched numerical variables were corrected as follows:

* The ‘?’ character was replaced, using the pandas replace method, with the numpy nan. This is a floating-point representation of ‘Not a Number’.
* The floating-point variable was subsequently converted to the Int64 data type because the data represented by the variable are discrete quantitative.
* Results of the correction are shown in Figure 2.

A screenshot of a computer code

Description automatically generated

Figure 2 - correction of mismatched data types

The mismatched qualitative data were corrected as follows:

* These values were imported as strings and were ordered as nominal variables.
* A list of the desired order for each variable was defined.
* The fields were converted to a pandas categorical data type and ordered using the ordering list was passed as an argument.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Task 2 – Data exploration

* Visualize Distributions: Use histograms or box plots to visually inspect the distribution of variables.
* Identify Missing Values, Outliers, or Errors –
  + Missing Values can be identified using the is.nan and summarising with the sum() function
  + Outliers can be detected using statistical methods such as the Z-score or the IQR (Interquartile Range) method.
  + Errors may arise from invalid entries in categorical variables (e.g., wrong labels or mixed types). Check for unusual or unexpected values.

There are two types of variables in the dataset to analyse for missing values, outliers, or errors in the data, and to detect skewness.

Skewness was used to show the distribution of quantitative variables. In this cases, the quantitative variables of interest are

num\_lab\_procedures

num\_procedures

num\_medications

number\_outpatient

number\_emergency

number\_inpatient

and number\_diagnoses

The qualitative variables of interest are:

These were analysed using frequency distributions of the variables using histograms.

**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.