**Assessment 1 – report**

**Task 1 - Examination of the data types.**

* 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.
* There are two types of data mismatch. First, pandas imports string data, and mixed data types as the Object Dtype. Second, where some of the mixed data types comprise numerical and string value.
* In the first mismatched, a function was written to convert all Object Dtypes to the string type.
* In the second mismatched Dtpyes, the number\_outpatient, number\_inpatient, number\_emergency variables were loaded as string types due to the presence of a ‘?’ character to denote missing/unknown values.
* 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.
* The DataFrame of the corrected mismatches is shown in Figure 3.

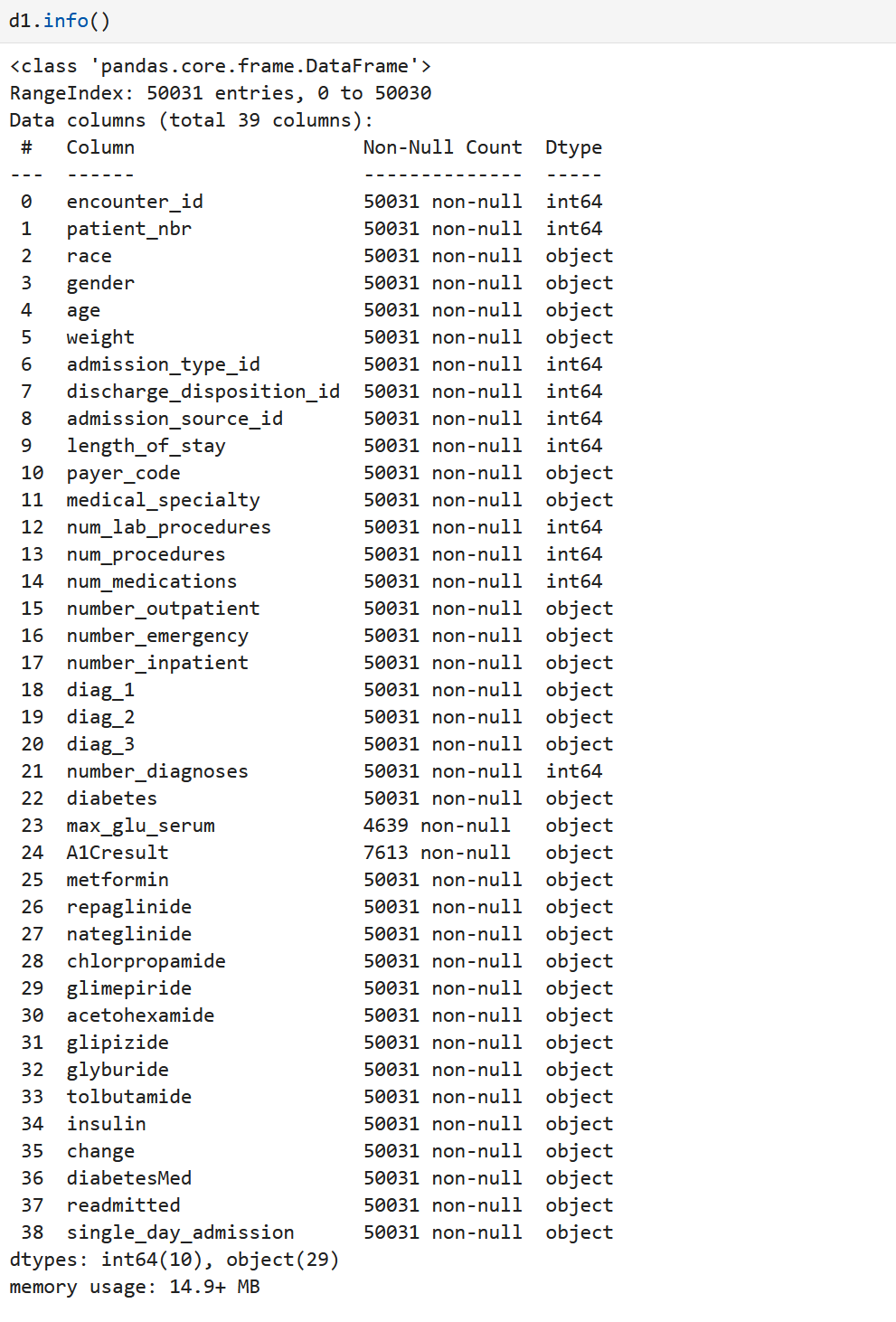


Figure 1 - Output of the pandas info() method on the imported dataset

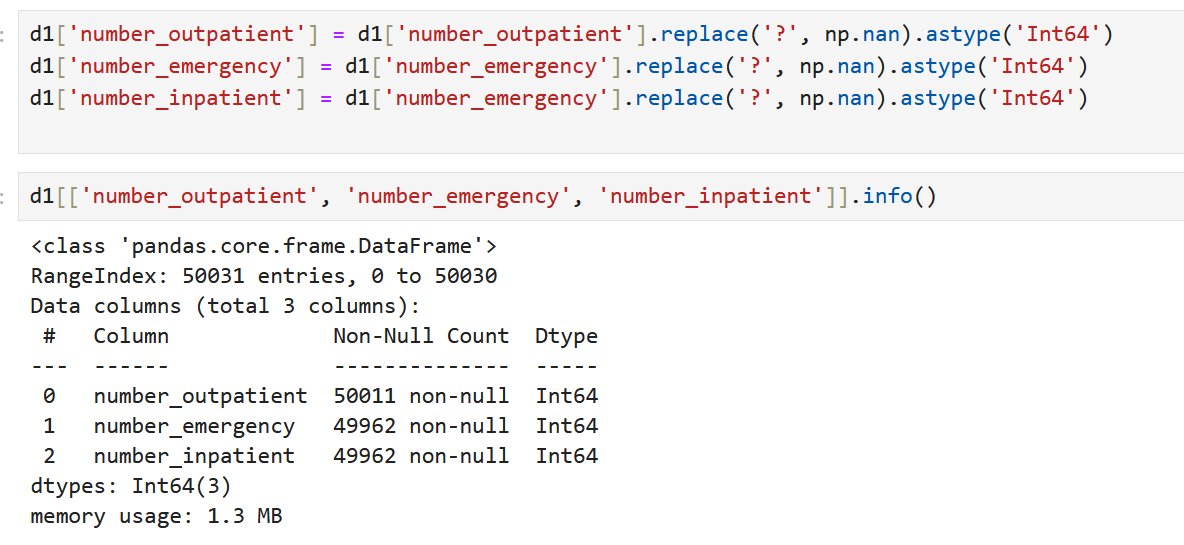


Figure 2 - Correction of mismatched data types

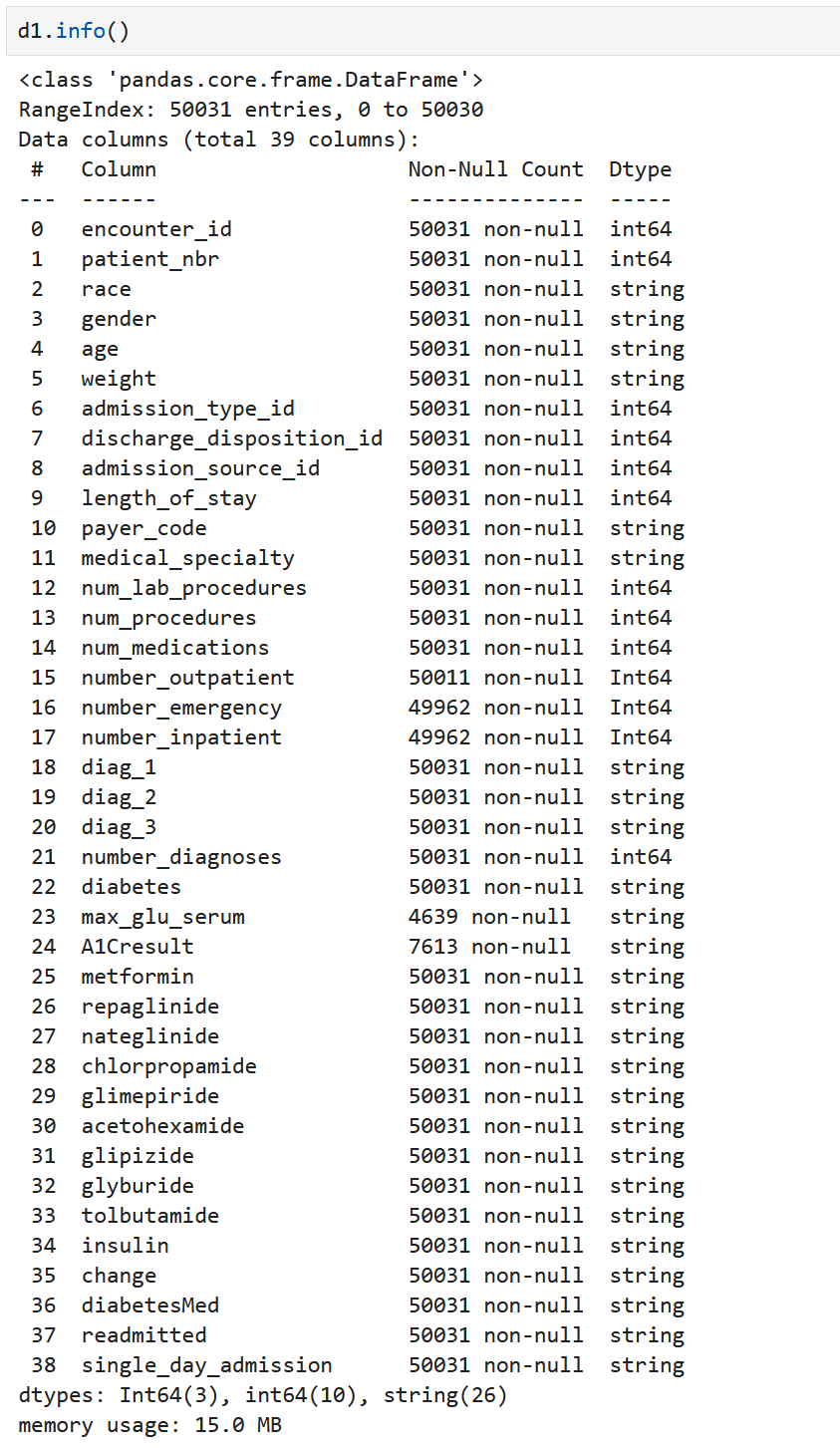


Figure 3 - Output of pandas info() after correction of mismatched data types

**Task 2 – Data exploration of the diabetes dataset.**

**Part 1 - Identify and report the skewness present in the variables**

Skewness is a statistical measure that describes the asymmetry of the distribution of quantitative numerical values in a dataset. In the case of our dataset, this can be measured on the meaningful integer values. Those variables are num\_lab\_procedures, num\_procedures, num\_medications, number\_outpatient, number\_emergency, number\_inpatient, and number\_diagnoses. The encounter\_id is an identifier variable and not required for analysis. The admission\_type\_id, discharge\_disposition\_id, and admission\_source\_id integervalues were not analused for skewness as these represent references to categorial variables.

A note about the patient\_nbr variable. This was not included at this stage for skewness as it can be excluded as an identifier variable. However, a small number of patients have multiple encounters that has its own frequency distribution. The variable was explored further as part of task 3.

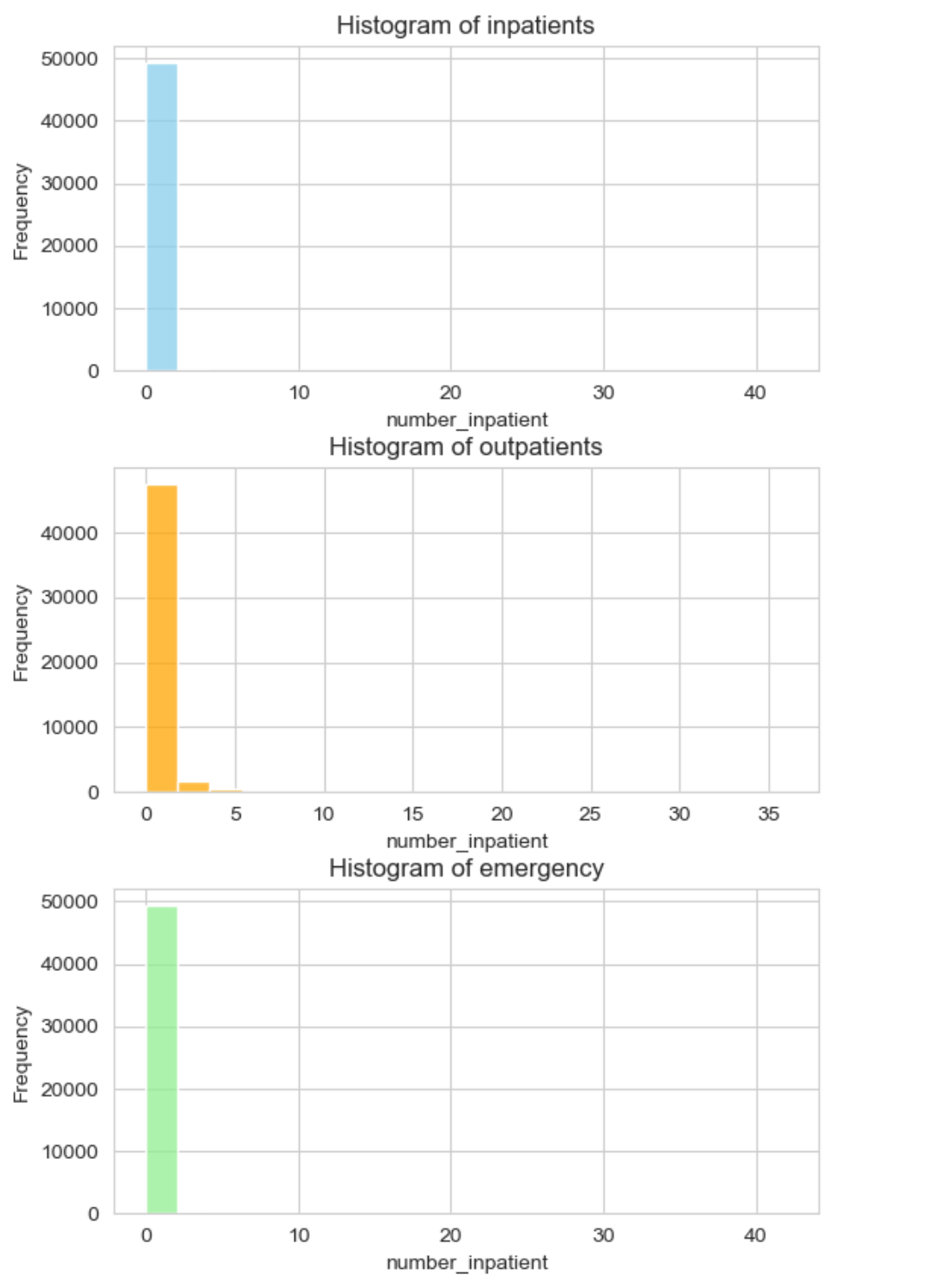
**Frequency distribution of the quantitative variables.**

* 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.

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.