**Assessment 1 – report**

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

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.

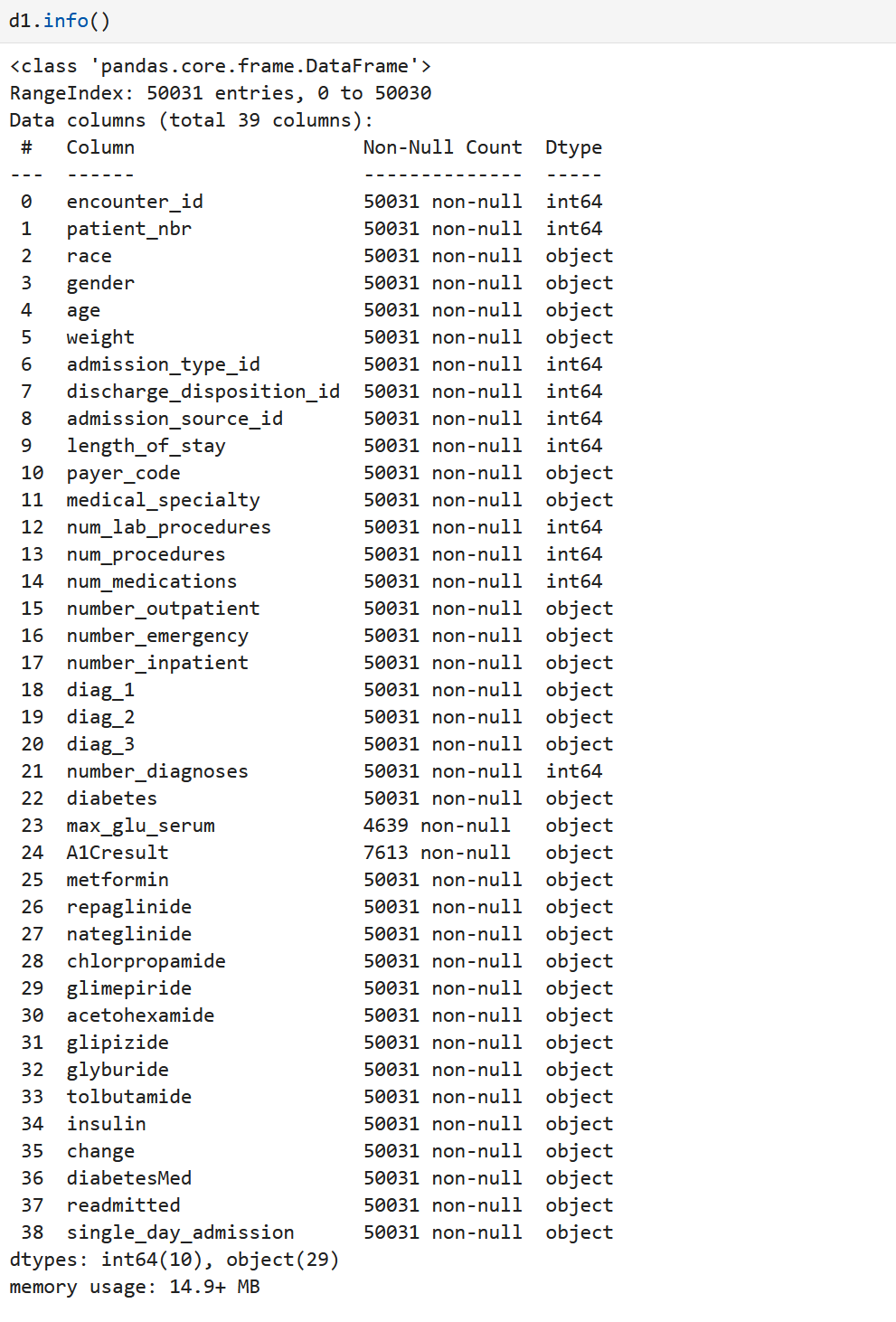


Figure 1 - Output of the pandas info() method on the imported dataset

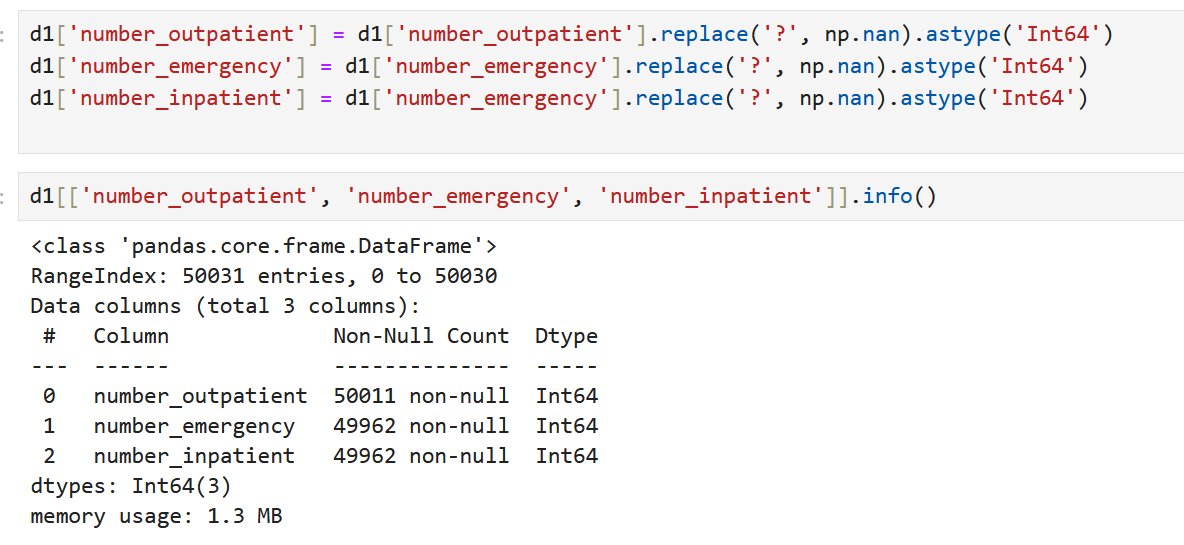


Figure 2 – Code to perform the correction of mismatched data types

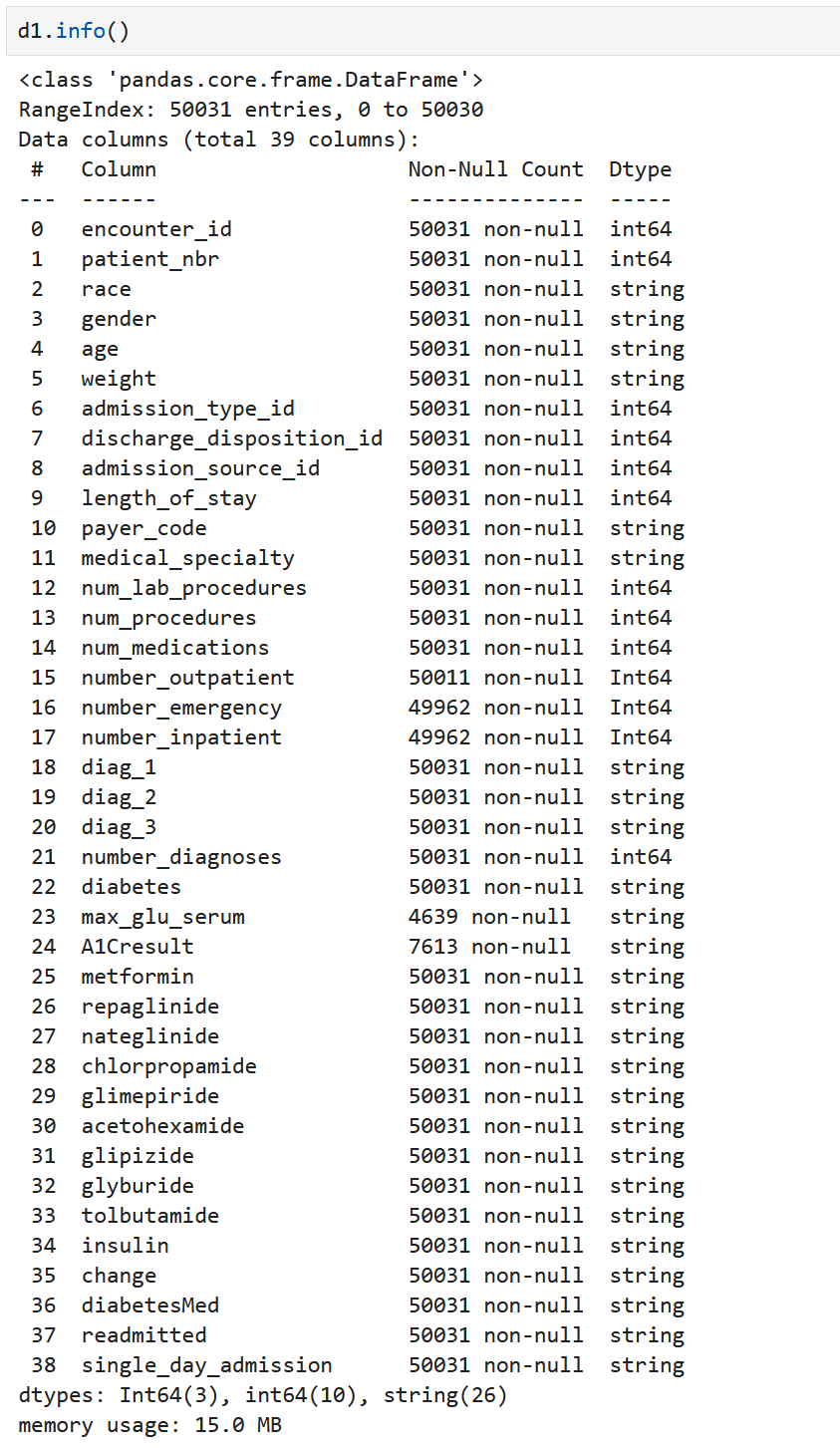


Figure 3 - Output of pandas info() after correction of mismatched data types

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

Using suitable statistical measures and functions, and visualisation plots as well to

* Identify and report the skewness present in the variables.
* Identify missing values, outliers, or errors in the data.
* List the variables with the identified problems.

**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 length\_of\_stay, 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 analysed 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.**

The skewness was determined by visual inspection of the frequency distribution of the quantitative variables. Also, values of skewness were reported using the numpy skewness method.

The quantitative variables very separated into three distinct groups to reflect different characteristics of the patients, their medical history, and encounters.

|  |  |
| --- | --- |
| Group | Variables include |
| Relating to prior healthcare exposure | Number\_inpatient, number\_outpatient, number\_emergency |
| Relating to procedures/medications during the encounters | Num\_procedures, num\_lab\_procedures, num\_medications |
| Relating to comorbidity and time in hospital | Num\_diagnoses, length\_of\_stay |

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 and boxplots using the code in Figure 4.

A screenshot of a computer code

Description automatically generated

Figure - code to produce the histograms and boxplots

A group of graphs with text

Description automatically generated with medium confidence

Figure - histograms and boxplots showing distributions of prior inpatient, outpatient, and emergency department visits

A close-up of a computer screen

Description automatically generated

Figure - Code block to produce the table of the percentage of encounters with number of visits per visit type

A screenshot of a graph

Description automatically generated

The table shows that for each visit type, the majority of patients had no prior visits to inpatient, outpatient, or emergency departments in the preceding year. This means that the majority of records for these three variables are outliers.

We tested this using the interquartile range (IQR) method to calculate the number and percentage of outliers for each variable. This method reported the number of values that either less than the 25th percentile multiplied by 1.5 times the IQR, or more than the 75th percentile times 1.5 times the IQR.

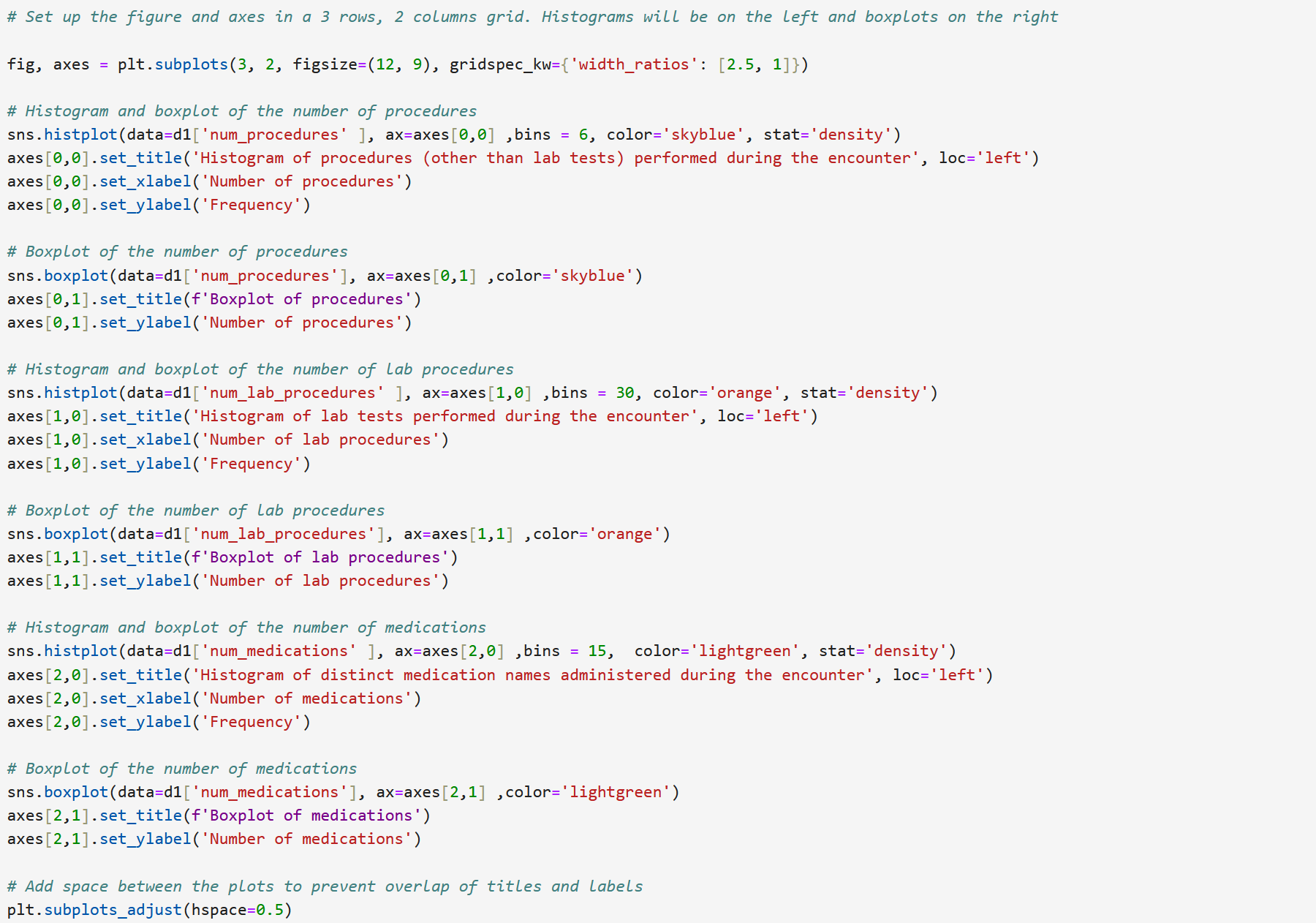
A function was written which incorporated the pandas describe() to calculate the number of outliers and percentage outliers in columns.

A screenshot of a computer code

Description automatically generated

A screenshot of a number

Description automatically generated



A group of graphs and diagrams

Description automatically generated with medium confidence

A screenshot of a computer code

Description automatically generated

A group of graphs and diagrams

Description automatically generated with medium confidence

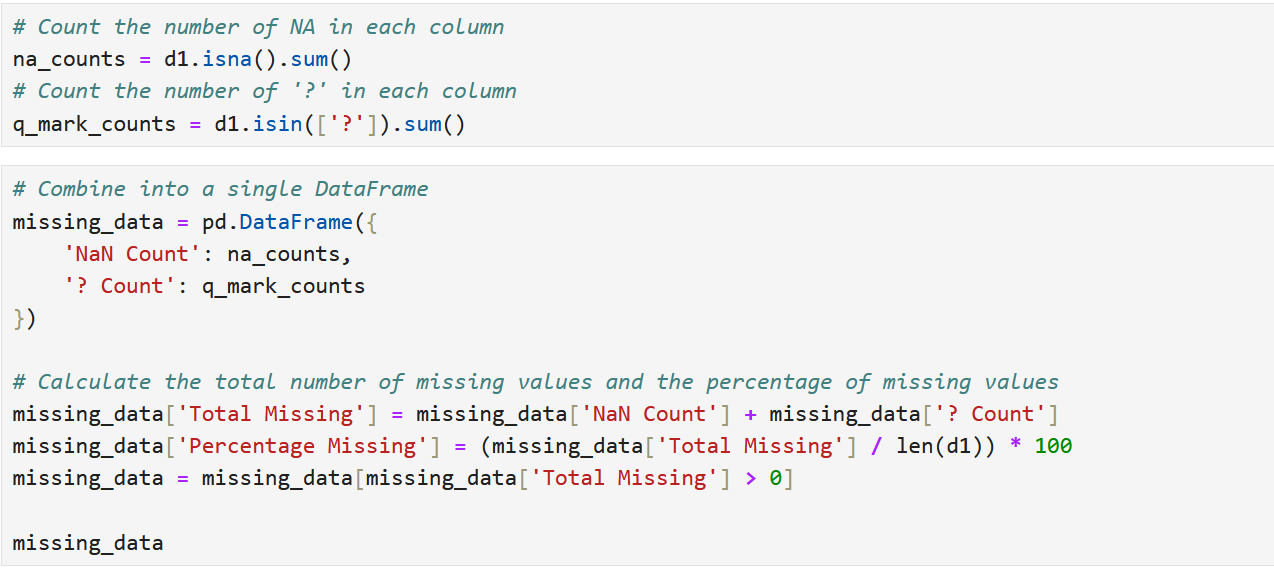
**Part 2 - Finding missing data and outliers**

Each variable was assessed for the number of unique values using the value\_counts() method with the drop\_na=False parameter. The output of the value\_count() method calls for selected variables is shown in figure X

A screenshot of a computer

Description automatically generated

It is evident that the ‘?’ character is used in the dataset to represent missing values. The data in this data set comprise strings and integer values. To calculate the number of missing values to different approaches were used. For integer values the sum of NaN values was calculated. For string variables, the number of ‘?’ characters were calculated and the the results with total values combined. These are presented in Table X



A screenshot of a graph

Description automatically generated

**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.