**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.

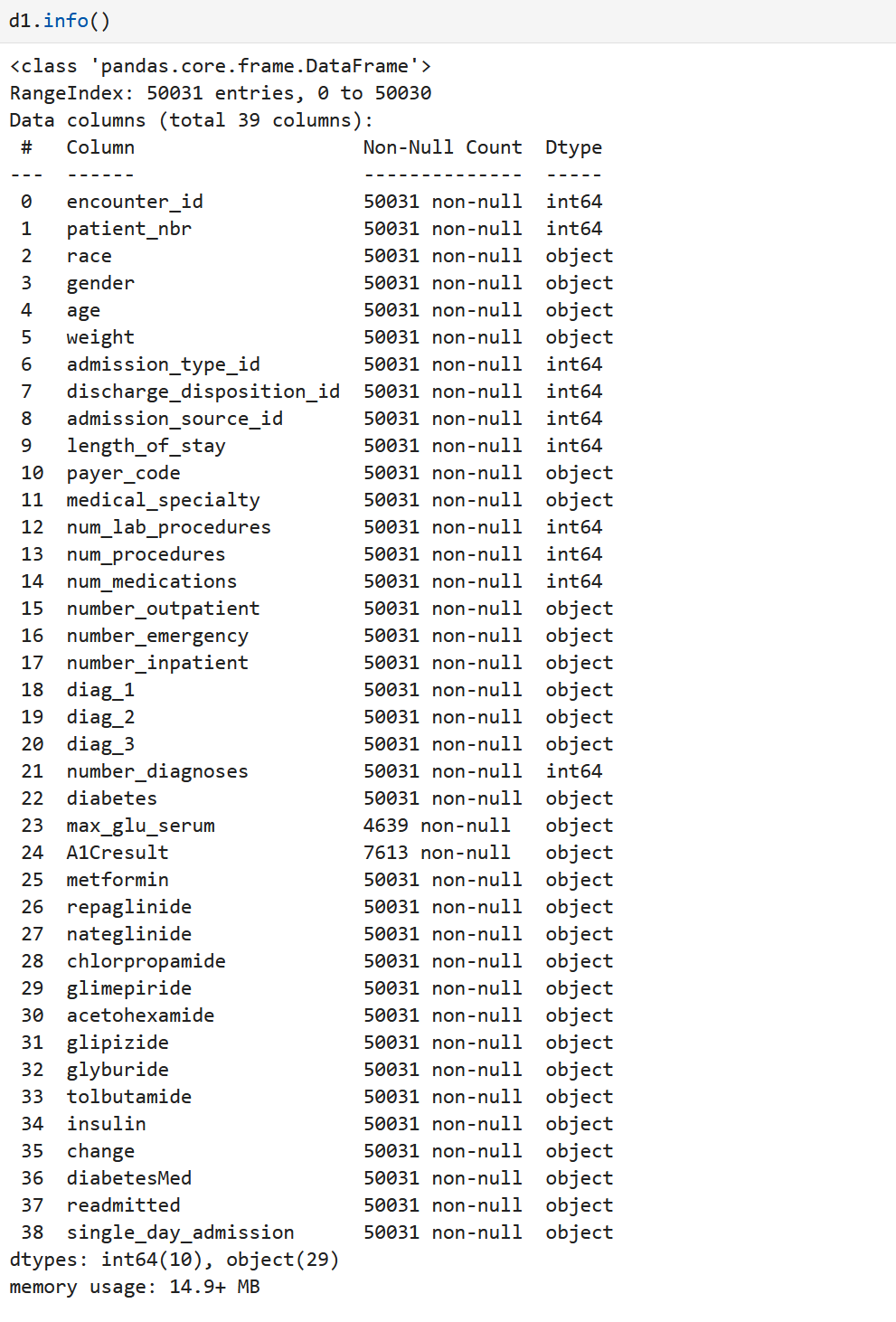


Figure 1 - Output of the pandas info() method on the imported dataset

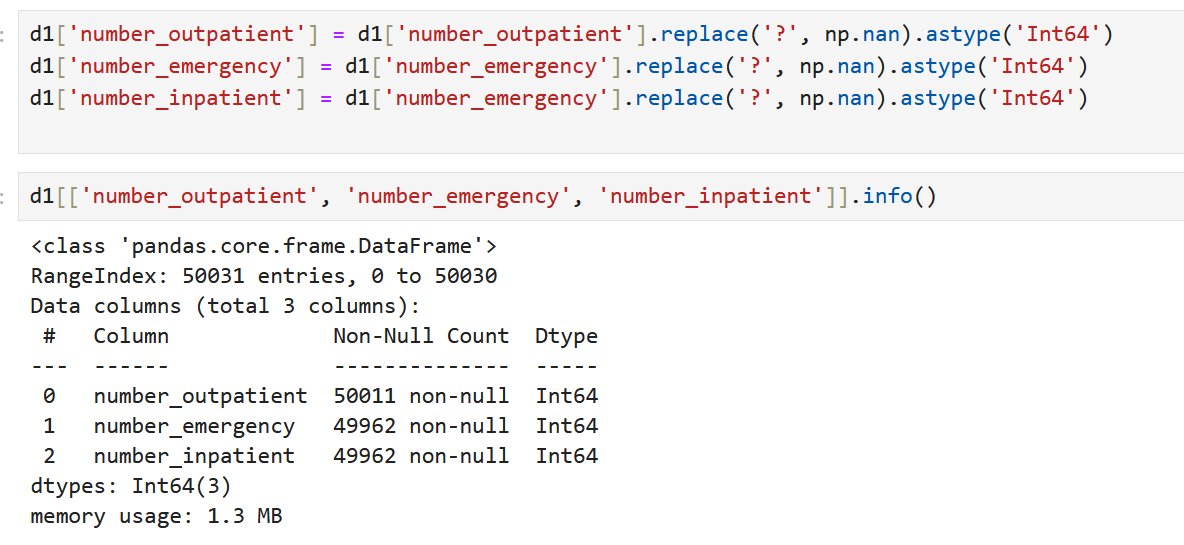


Figure 2 – Code to perform the correction of mismatched data types

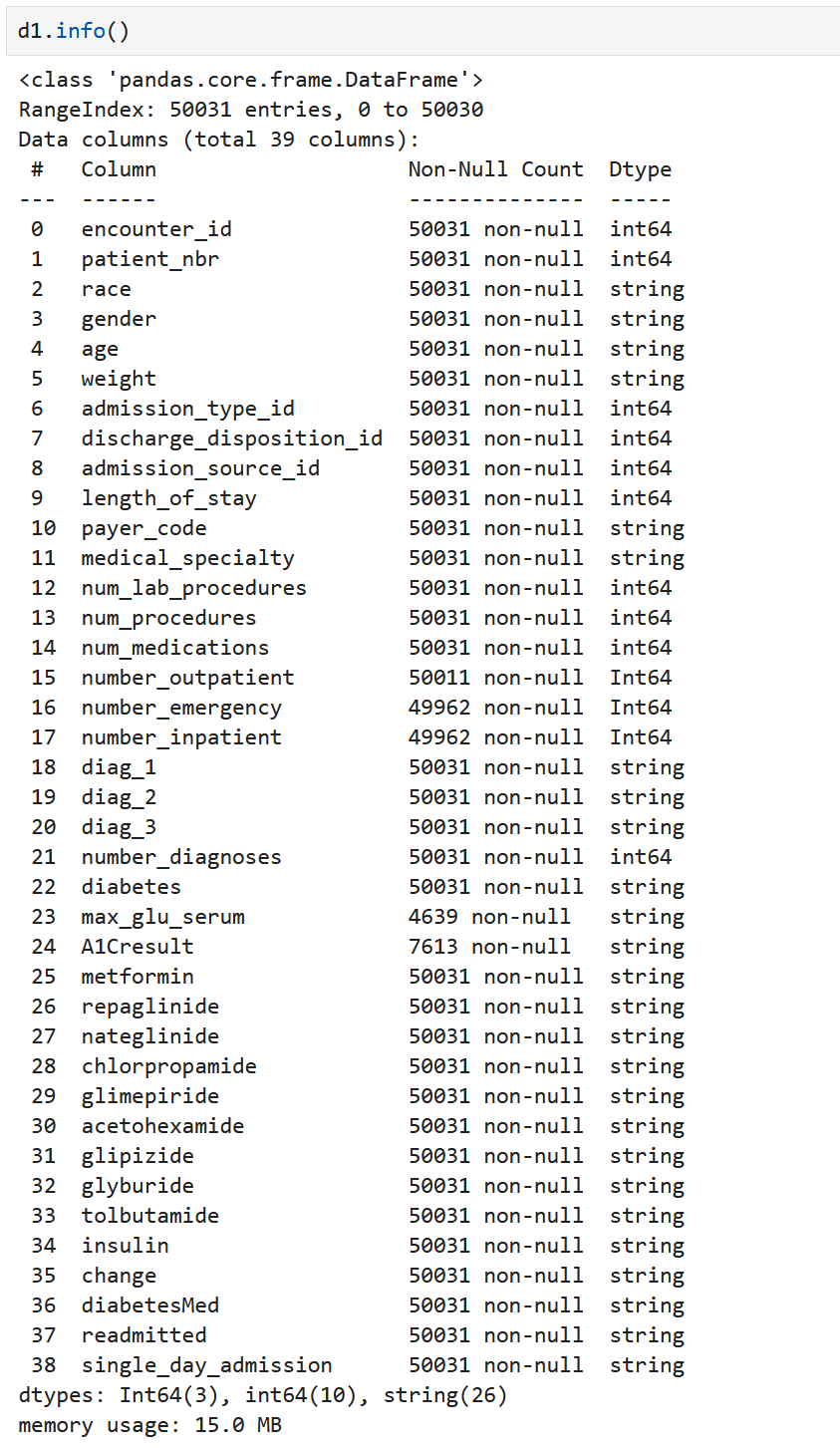


Figure 3 - Output of pandas info() after correction of mismatched data types

**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.

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, most patients had no prior visits to inpatient, outpatient, or emergency departments in the preceding year. This means that many records for these three variables are outliers. All variables in this group are strongly positively skewed.

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.

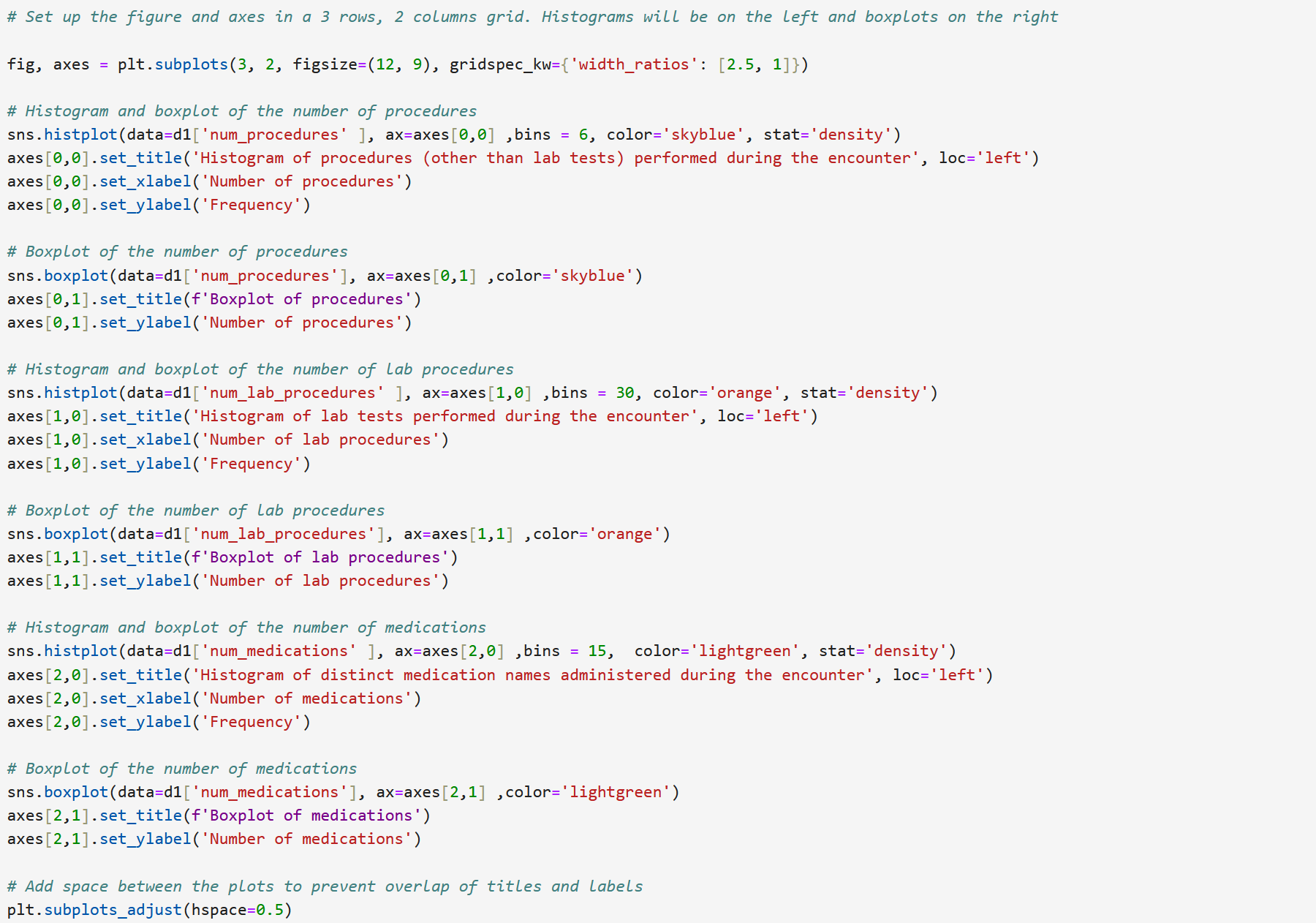
A screenshot of a computer code

Description automatically generated

A screenshot of a number

Description automatically generated

These outliers are a result of the nature of the data. The data comprises zero values therefore values above can quickly fall into the outlier category as the value of visits increases. outliers for this group of variable will require careful consideration in terms of handling the abundance of zeroes.



A group of graphs and diagrams

Description automatically generated with medium confidence

These outliers should be considered differently than the outliers from patient visits. For example, while the number procedures show 6 as a statistical outlier based on the IQR, it is substantively not an outlier because it is the range of a small number of discrete procedures variable.

The distributions of the procedures/medications variables are all positively skewed and the handling of outliers for this group of variable will require careful consideration for handling the values in the long tails.

A screenshot of a computer code

Description automatically generated

A group of graphs and diagrams

Description automatically generated with medium confidence

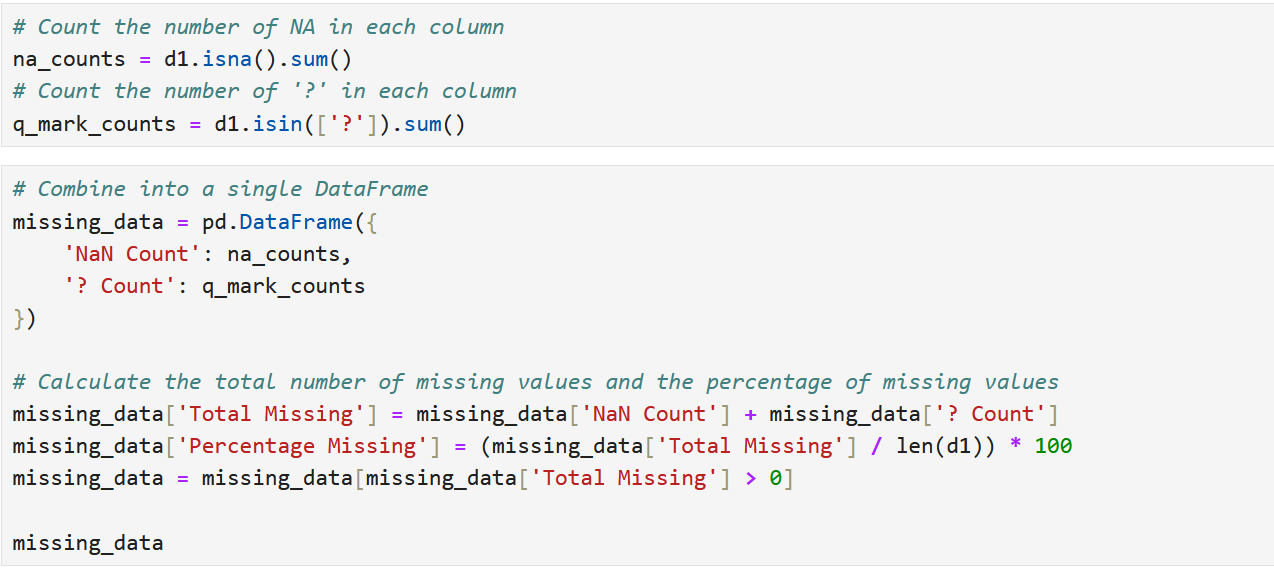
**Part 2 - Finding missing data and outliers**

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

These methods have shown several issues in the data that require correction before data mining can commence. In addition to the issues identified with quantitative variables, some issues were identified with qualitative variables when reviewed with the value\_counts() method. For example, the age and weight variables, in addition to having missing values, are loaded as nominal values. These were converted to ordinal values by assigning a order with the pd.Categorial() function. This will provide options for either one-hot or ordinal encoding depending on the data mining approach chosen for the analysis.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The diagnosis categories (diag\_1, diag\_2, and diag\_3) represent ICD9 codes of primary, secondary, a tertiary diagnoses of cases. The variables each have more than 600 unique values. This would be impractical for a machine learning model so these will require grouping.

Similarly, medical speciality has 68 unique values and 35% of values are missing. This variable will be recoded into higher level specialities e.g. internal medicine, surgery, general practice, oncology, paediatric etc.

List of values with identified problems:

|  |  |  |
| --- | --- | --- |
| Variable | Problem | Proposed correction |
| encounter\_id | None | Drop – not required for data mining |
| patient\_nbr | None |  |
| race | Missing values | Impute by mode imputation |
| gender | 1 missing value | Impute by mode imputation |
| age | Loaded as nominal binned variable | Convert to ordinal variable |
| weight | 96% missing | Drop – majority of values are missing |
| admission\_type\_id | Lookup reference variable | No action required |
| discharge\_disposition\_id | Lookup reference variable | No action required |
| admission\_source\_id | Lookup reference variable | No action required |
| Length\_of\_stay | None | No action required |
| payer\_code | 65% Missing | Drop – too many missing values. |
| medical\_specialty | Multiple unique categories and 35% missing | Recode to higher level categories. |
| num\_lab\_procedures | None | No action required |
| num\_procedures | None | No action required |
| num\_medications | None | No action required |
| number\_outpatient | Small number of missing values.  Strongly zero inflated. | Different options for transformation are available, depending on model choice. These are:   * Square root transformation * Log transformation with offset for zero values. * Convert the three variables into a composite health\_history variable |
| number\_emergency | Small number of missing values.  Strongly zero inflated. |
| number\_inpatient | Small number of missing values.  Strongly zero inflated. |
| diag\_1 | Many unique values | Recode to top level ICD9 categories |
| diag\_2 | Many unique values | Consider ignoring as could be on limited value compared to primary diagnosis |
| diag\_3 | Many unique values | Consider ignoring as could be on limited value compared to primary diagnosis |
| number\_diagnoses | None | No action required |
| diabetes | None | No action required |
| max\_glu\_serum | Many missing values | Drop – too many missing |
| A1Cresult | Many missing values | Drop – too many missing |
| metformin | None | No action required |
| repaglinide | None | No action required |
| nateglinide | None | No action required |
| chlorpropamide | Missing value | Correct missing value |
| glimepiride | None | No action required |
| acetohexamide | Single value only | Drop – no variability. Contributes no information. |
| glipizide | None | No action required |
| glyburide | None | No action required |
| tolbutamide | Single value only | Drop – no variability. Contributes no information. |
| insulin | None | No action required |
| change | None | No action required |
| diabetesMed | None | No action required |
| Readmitted | None | No action required |
| single\_day\_admission | None | No action required |

**Task 3 – Data preparation**

This will be the bulk of the work

* Summarise findings
* Undertake data preparation – this will be, for example,
  + recategorizing variables e.g. diagnosis codes, medical speciality,
  + normalising/standardizing variables with outliers e.g. admission days,

**Task 4 - Selection of data mining task and feature selection**

* Identify correlation between variables (use a correlation matrix)
* Examine relationship between diabetes and diabetesMed – this will be strong. DiabetesMed will be redundant.
* Exclude pointless variables e.g. weight, payercode, max\_glu\_serum, AC!result etc…
* Identify mining task – this will likely be a classification algorithm using supervised machine learning. For example,
  + are there characteristics of the patient encounters that can be used to predict hospital length of stay. Or,
  + are there characteristics of the patient encounters that can be used to predict mortality as an outcome?
* What variables should be included and why?