**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.

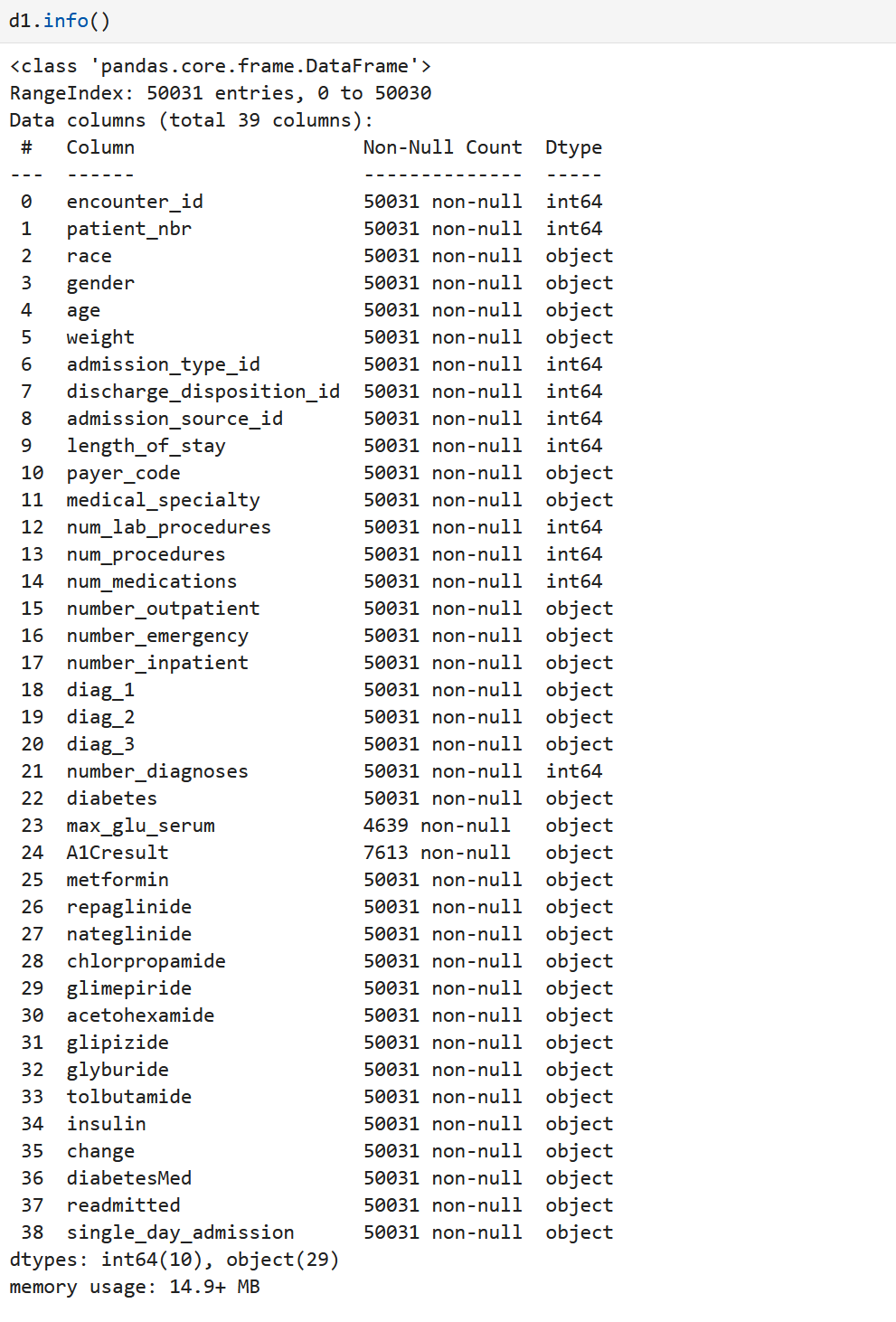


Figure 1 - Output of the pandas info() method on the imported dataset

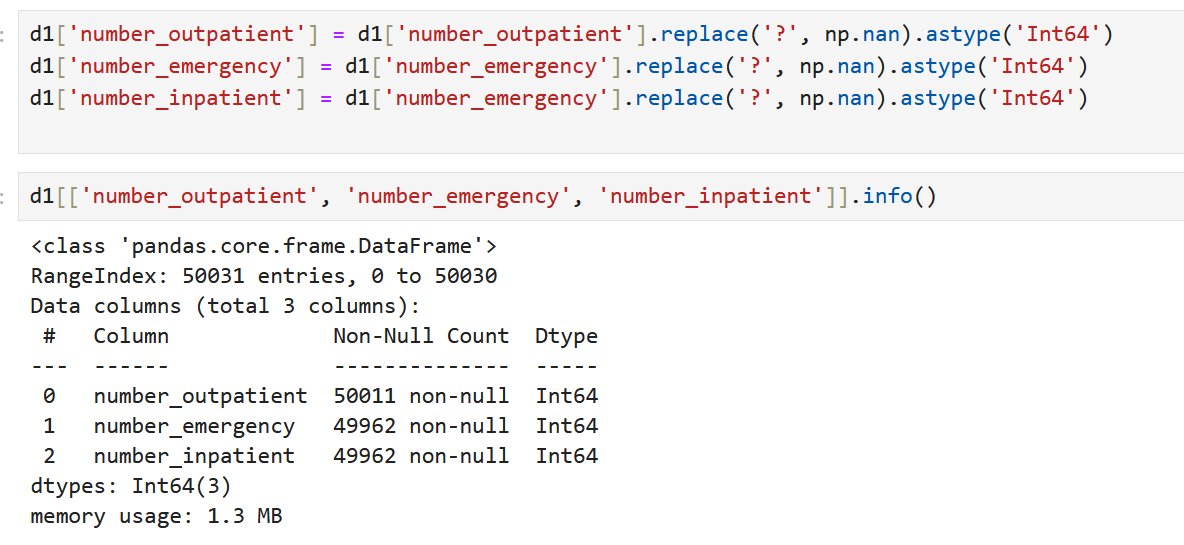


Figure 2 – Code to perform the correction of mismatched data types

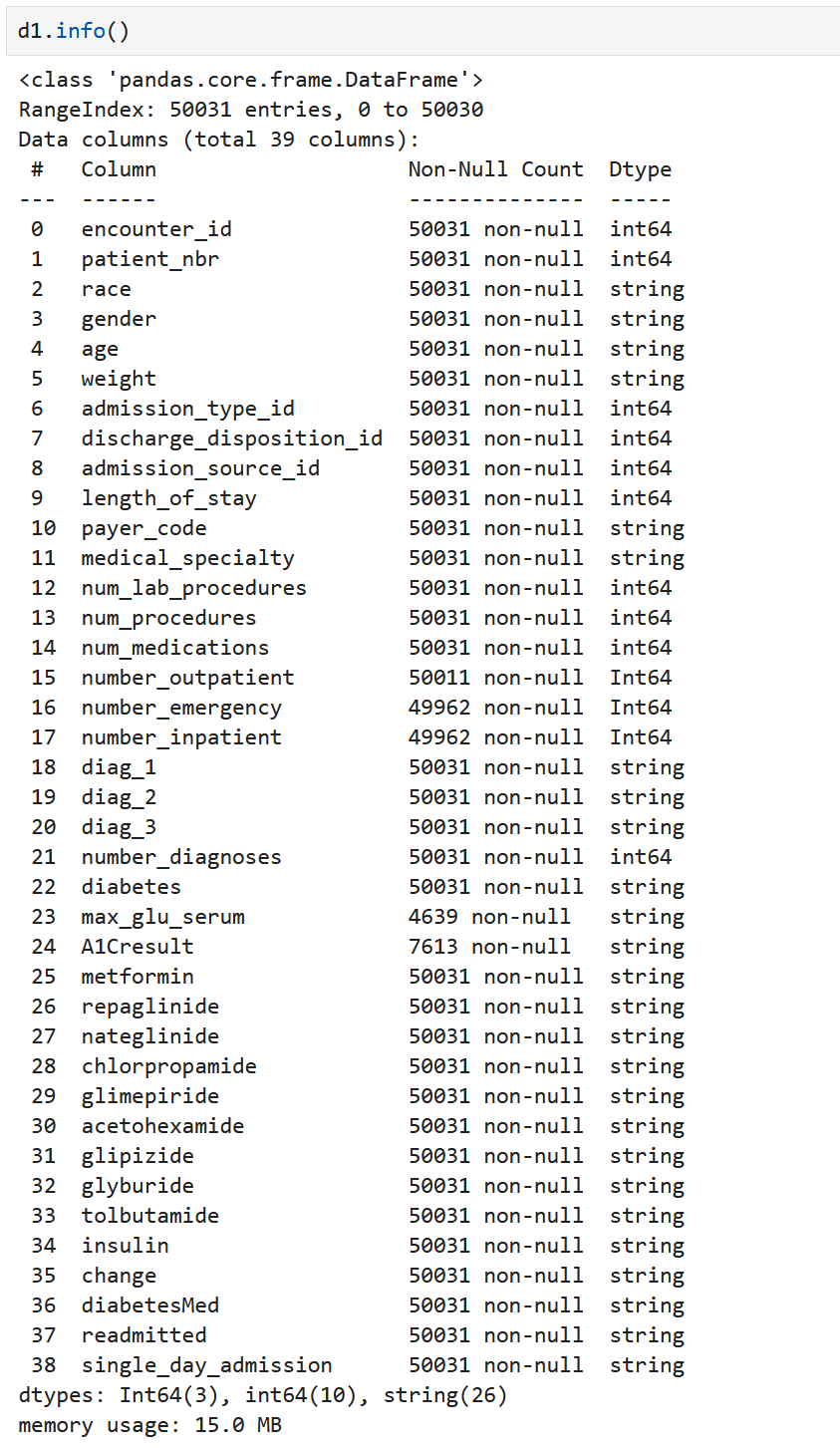


Figure 3 - Output of pandas info() after correction of mismatched data types

**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

The histograms show that the number of inpatient, outpatient, and emergency visits are right-skewed, with most patients having few visits. The boxplots show that there are many outliers in each of the variables. This is expected given the right-skewed distribution of the data.

This can be corroborated by calculating the skewness of the variables.

A screen shot of a computer

Description automatically generated

The skewness values are positive, confirming that the data are right-skewed which is consistent with the histograms and boxplots.

The following code shows the zero-inflated nature of these group1 numerical data types.

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.

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

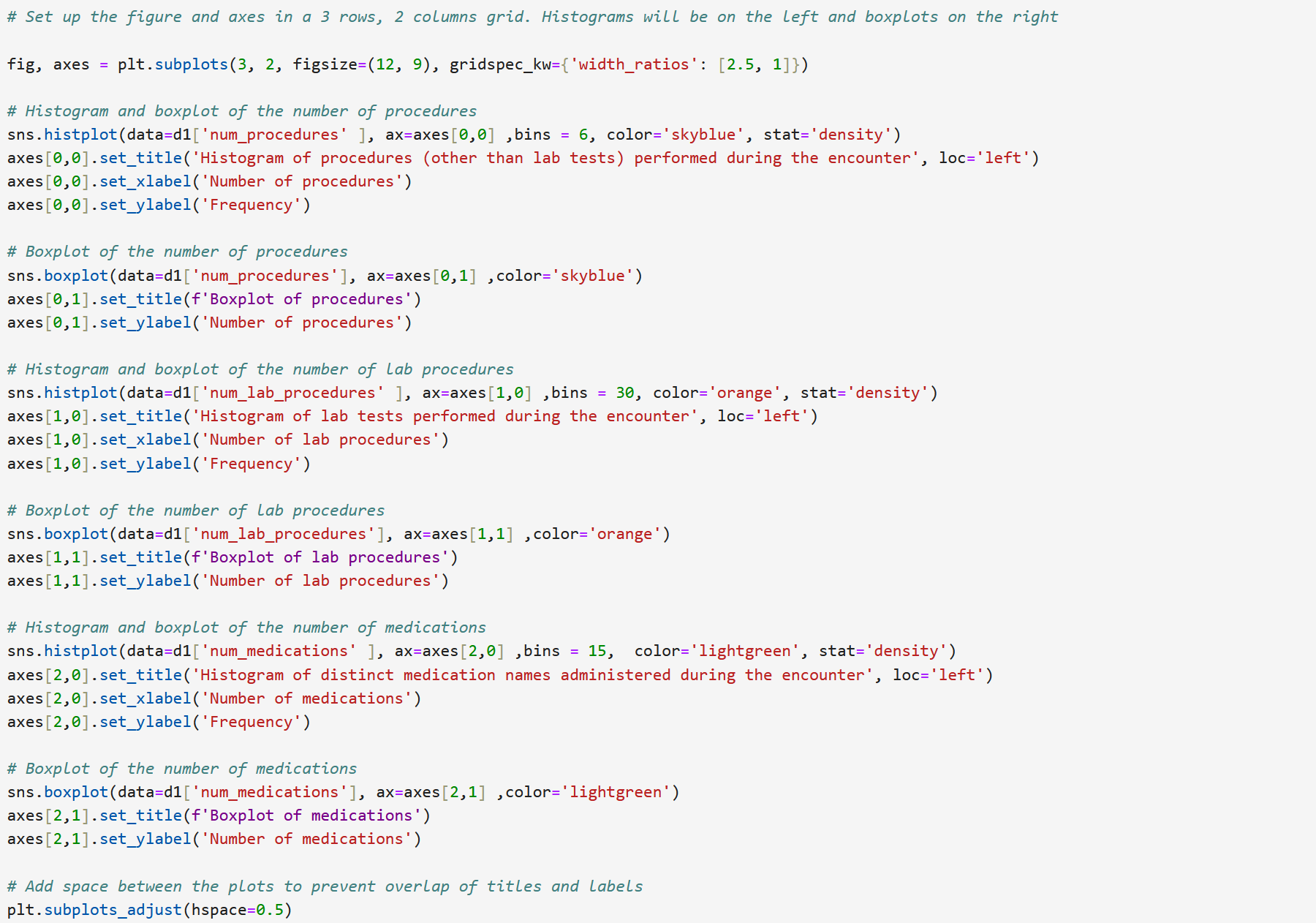
Description automatically generated

A screenshot of a number

Description automatically generated

These outliers are a result of the data being strongly zero-inflated therefore the IQR and outlier upper bound are low values.

The same process was repeated for the second group of variables representing number of procedures, lab tests, and medications.



A group of graphs and diagrams

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

The histograms show that the number of procedures and medications are positively skewed. With a smaller number of patients having large numbers of procedures and medications.

The number of lab procedures is more evenly interesting because it shows what appears to be a long tail, in the presence of several zero values. The skewness variable indicates that this variable is in fact slightly negatively skewed.

A screenshot of a computer

Description automatically generated

The boxplots show that there are many outliers in each of the variables. This needs to be interpreted in the context of the data. For example, the number of procedures and medications, while exceeding the IQR, may not be considered outliers in a medical context.

Also, the number of outliers in the lab procedures is comparatively low, but this is due to the large number of zero values in the data. This is likely due to the fact that not all patients require lab procedures during their encounter.

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

The histograms show that the number of diagnoses are negatively skewed whereas the length of stay is positively skewed. The boxplots show that there are no outliers for the number of diagnoses, but there are many outliers for the length of stay.

A screenshot of a computer

Description automatically generated

These are corroborated by the skewness values which are negative for the number of diagnoses and positive for the length of stay.

**A screenshot of a computer program

Description automatically generated**

**Part 2 - Identify and show variables with missing data**

Missing data are represented by '?' in the dataset. For integer values that have already been replaced with 'NaN'. This code allows us to detect them using the isna() method for NaNs, and the isin() method to search for '?' characters in qualitative variables.

The data in the admission\_type\_id, discharge\_disposition\_id, and admission\_source\_id columns are coded using reference codes. The missing values are represented by codes that are not in the reference list. These codes need to be identified and mapped to the correct missing value co

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

**Task 3 – Data preparation**

Summary here

* 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,

List of values with identified problems:

|  |  |  |
| --- | --- | --- |
| Variable | Problem | Proposed correction |
| encounter\_id | None | Drop – not required for data mining |
| patient\_nbr | Multiple records | Drop |
| race | Missing values | Impute by random value 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 | None (lookup value) | Identify and correct missing values.  Recode variable to reduce the number of features. |
| discharge\_disposition\_id | Lookup reference variable | Identify and correct missing values.  Recode variable to reduce the number of features. |
| admission\_source\_id | Lookup reference variable | Identify and correct missing values.  Recode variable to reduce the number of features. |
| 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 | The variables have a small number of missing values.  The variables are strongly zero inflated. | Replace NaN values with median value to limit bias. |
| number\_emergency |
| number\_inpatient |
| diag\_1 | Many unique values | Recode to top level ICD9 categories |
| diag\_2 | Many unique values | Recode to top level ICD9 categories |
| diag\_3 | Many unique values | Recode to top level ICD9 categories |
| 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 |  | These values will be transformed with one-hot encoding, except acetohexamide and tolbutamide. The latter two have single values and as such no variability therefore would not contribute meaningful information to a machine learning algorithm.  Chlorpropamide has a missing value that will be imputed. |
| repaglinide |
| nateglinide |
| chlorpropamide |
| glimepiride |
| acetohexamide |
| glipizide |
| glyburide |
| tolbutamide |
| insulin |
| change | None | No action required |
| diabetesMed | Possibly redundant | Consider dropping |
| readmitted | None | No action required |
| single\_day\_admission | None | No action required |

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.

The variables admission\_type\_id, discharge\_disposition\_id, and admission\_source\_id can be recategorized to reduce the number of values.

The admission\_type\_id can be condensed to fewer categories as follows:

* Emergency, urgent, and trauma center can be combined into the Emergency category based on their acuity
* Not Available, NULL, and Not Mapped can be combined into the Not available category based on the lack of mapping, or unavailability
* Newborn can be left as a separate category

The discharge\_disposition\_id can be condensed to four distinct categories that capture the concepts of discharged home; transferred for ongoing care; deceased; and unknown.

* Discharged to home can be categorised as 'Discharged home'
* Examples of discharged/transferred include another short-term hospital, SNF, ICF etc can be categorised as 'Ongong care'. This is because they are all forms of additional care in the current facility or through another facility.
* Any variable with Expired will be recoded to 'Deceased'
* All ‘Not mapped’, NULL, and Unknown will all be recoded to 'Unknown'

The admission\_source\_id can be condensed to fewer categories as follows:

* Referral from another provider e.g. a physician or clinic.
* Transferred in from another facility or health agency.
* Emergency room admission.
* Not Available

On inspection of the variables, there were a small number that were related to childbirth/newborns and some that were court order. Given the small number (less than 10) these were considered as a referral.

**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?