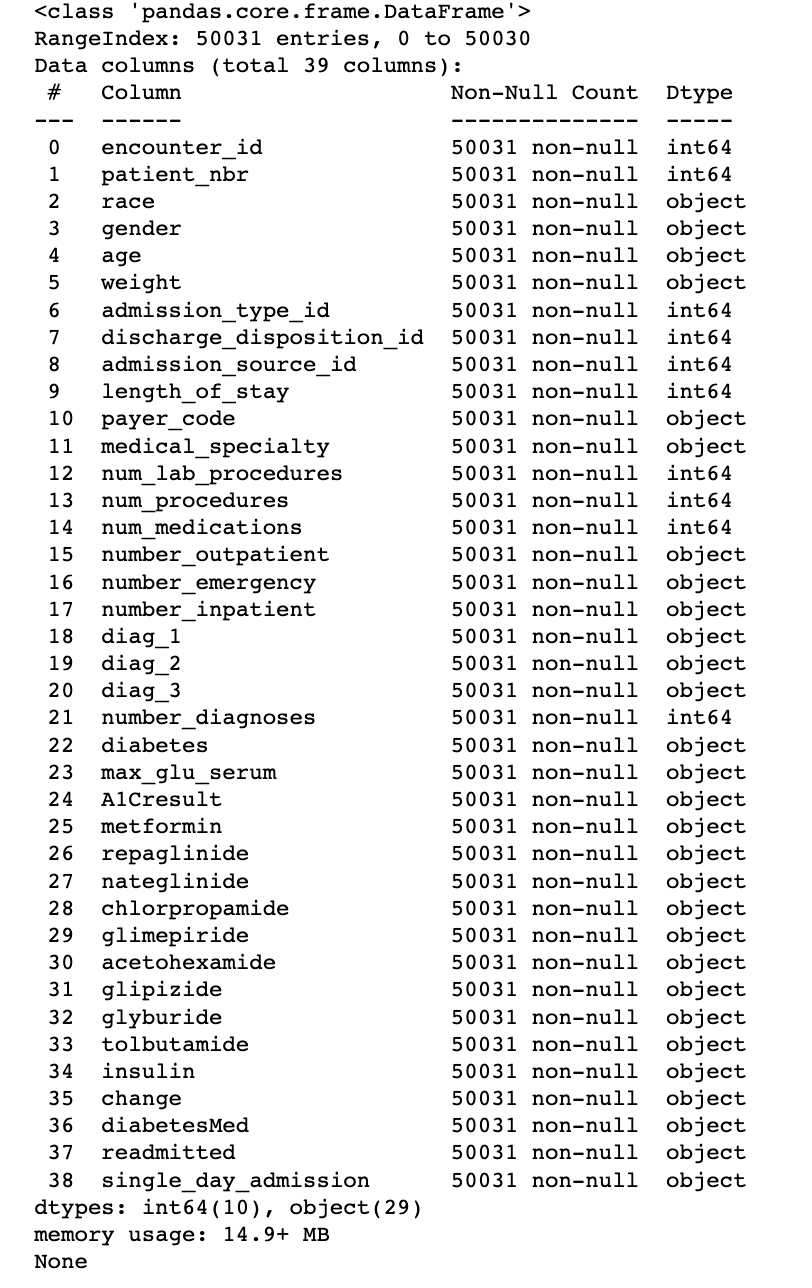
IFN509 Assignment 1

Renee Chand n1088373 , Melanie Packer n5736404, Phan Thao Nhi Nguyen n11232862

# Question 1: Variable Data Types

The dataset was perused and identified using the df.info() function. The following screenshot shows the Dtype of the 39 variables prior to pre-processing.



The identified mismatched data types are as following:

1. Number outpatient
2. Number inpatient
3. Number emergency
4. Diabetes
5. DiabetesMed
6. Change
7. Single day admission
8. Gender
9. Age
10. Race
11. Weight
12. Medical speciality
13. Maximum glucose serum
14. A1C result
15. Metformin
16. Repaglinide
17. Nateglinide
18. Chlorpropamide
19. Glimepiride
20. Acetohexamide
21. Glipizide
22. Glyburide
23. Tolbutamide

The following variables (number\_outpatient, number\_emergency and number\_inpatient) were converted to int64 as they are discrete in nature.

1. Number outpatient
2. Number inpatient
3. Number emergency

The following variables (diabetes, change, diabetesMed and single\_day\_admission) were converted to bool. These variables had only two potential options whereby bool was an appropriate data type choice. The correct identification of these variables are categorical.

1. Diabetes, with 1 being yes and 0 being no
2. DiabetesMed- with 1 being yes and 0 being no
3. Change- with 1 being change and 0 being no change
4. Single day admission- with 1 being admission and 0 being no admission

Gender was converted to binary data type using binary identification of 0 as female and 1 as male as these were the two genders identified in the dataset.

1. Gender

The following 17 variables were converted to categorical data type as these variables consist of grouped data.

1. Age
2. Race
3. Weight
4. Medical speciality
5. Maximum glucose serum
6. A1C result
7. Metformin
8. Repaglinide
9. Nateglinide
10. Chlorpropamide
11. Glimepiride
12. Acetohexamide
13. Glipizide
14. Glyburide
15. Tolbutamide
16. Insulin
17. Readmitted

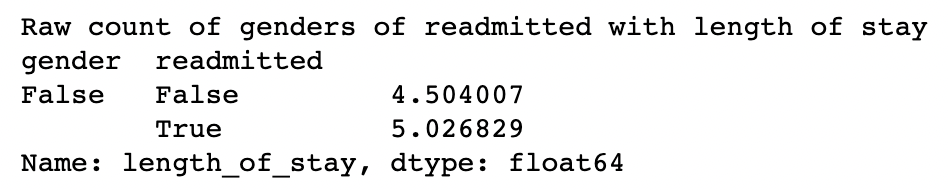
Age and weight variables were ordered for appropriate visualisation later, presented as an ordinal variable.

The appropriate screenshot displaying the correct data type of variables is shown below:

# Question 2: Exploration with Statistical Measures

1. The identified and reported skewness in the variables are described and shown below:
2. Length of stay
3. Number of laboratory procedures
4. Number of procedures
5. Number of medications
6. Number outpatient
7. Number emergency
8. Number inpatient
9. Number diagnoses
10. There were several inconsistencies, errors and missing values identified in the data. The errors have been identified and corresponding detail identified.
11. The following questions were answered using various Python packages:
12. The average length of stay in the hospital for a male patient who was readmitted in less than 30 days.

The data was grouped by gender and readmission. The conditions of male and less than 30 days readmission were filter, with the length of stay being the variable of interest. As the average length was asked, the mean was the appropriate function. The average length of stay in the hospital for a male patient who was readmitted in less than 30 days is 5.02 days or rounded to 5 days as length of stay is a discrete number. True is male gender and False is a female gender. The screenshot below verifies this.



1. The age group that has the highest number of encounters whose primary diagnosis is diabetes is

The number is

1. There were \_\_\_\_\_ encounters whose admission type is Emergency.

There were \_\_\_ Emergency encounters readmitted within 30 days.

1. The top-three race categories according to the number of readmission cases (including both less than or larger than 30 days) are

# Question 3: Exploration with Visualisation Plots

1. The distribution of the variables and appropriate data quality problems are:
2. To determine if a relationship exists between diabetes and diabetesMed,

If a relationship exists between the two variables, appropriate data modelling could include

1. The highly correlated variable pairs include:

The appropriate mining process for these variables include:

# Question 4: Data Presentation

1. The findings based on the data exploration include
2. The data preparation steps to address the data quality problems include
3. The data preparation is shown by a screenshot of the Python code below

The data quality problems have been identified from the output

# Question 5: Feature and Task Selection

1. The most suitable data mining task to be performed on this dataset is:

This can be justified due to

1. The appropriate variables to include in this data mining task would be \_\_\_\_\_ due to \_\_\_\_\_\_\_

The roles include \_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_

There may be derived variables that include \_\_\_\_\_\_

# General