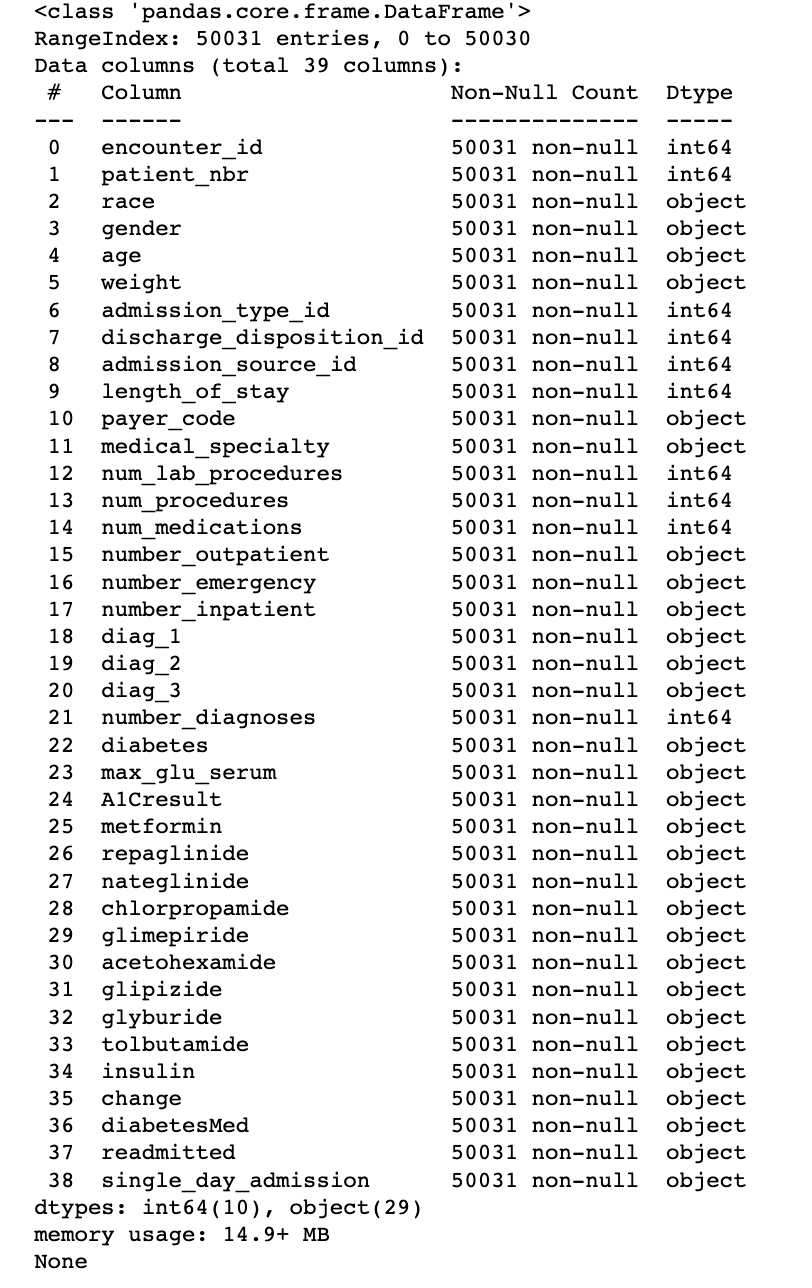
IFN509 Assignment 1

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

The correct mismatched data types have been identified and are described below.

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, gender and single\_day\_admission) were converted to bool (using the binary identification of 1/0). 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
2. DiabetesMed
3. Change
4. Single day admission
5. Gender

The appropriate screenshot/table 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:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

1. There were several inconsistencies, errors and missing values identified in the data. The errors have been identified and corresponding detail identified.
2. The following questions were answered using various Python packages:
3. The average length of stay in the hospital for a male patient who was readmitted in less than 30 days is
4. 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