IFN 509: Data Exploration and Mining

Assessment 1

Team Name: [ Bench-Push-100KG ]

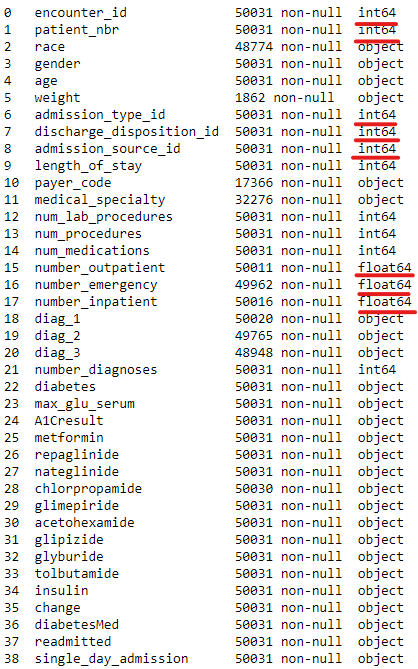
Group No. [ 9 ]

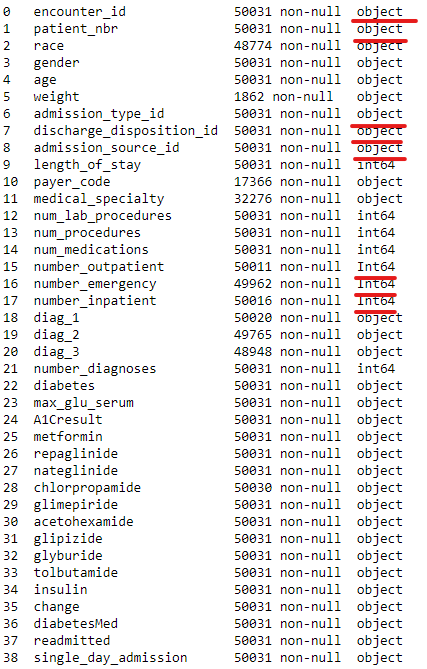
| Student Name | Student Id |
| --- | --- |
| Yuliya Manueva | N11235705 |
| Chih-Yu Chuang | N10952535 |
| Jordan Su | N11189983 |

|  | Student 1 | Student 2 | Student 3 |
| --- | --- | --- | --- |
| Student 1 | <100 %> | <100 %> | < 100%> |
| Student 2 | <100 %> | <100 %> | <100 %> |
| Student 3 | <100 %> | <100 %> | <100 %> |

Replace the % contribution with an appropriate number if it is not an equal contribution.

Task1





Task2

length\_of\_stay: Positively skewed

Mode:3 < Median:4 < Mean:4.5

num\_lab\_procedures:Negative skewed

Mean:42.3 < Median:43 < Mode:0/43

num\_procedures: Positively skewed

Mode:0 < Median:1 < Mean: 1.37

num\_medications: Positively skewed

Mode:12 < Median:14 < Mean15.28

number\_outpatient: Positively skewed

Mode: 0 <= Median: 0 < Mean: 0.22

number\_emergency: Positively skewed

Mode: 0 <= Median: 0 < Mean: 0.12

number\_inpatient: Positively skewed

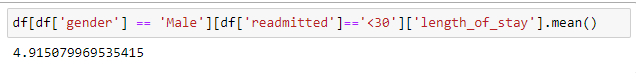
Mode: 0 <= Median: 0 < Mean: 0.58

number\_diagnoses: Negatively skewed

Mean:6.89 < Median:7 < Mode:9

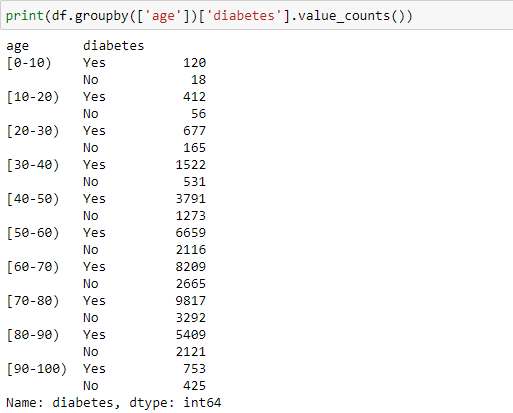
**a. What is the average length of stay in the hospital of a male patient who was readmitted in less than 30 days?**

Answer: 4.915079969535415 dys



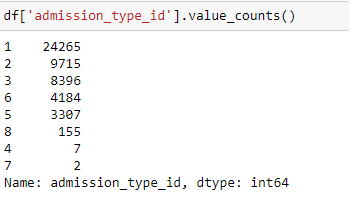
**b.Which age group has the highest number of encounters whose primary diagnosis is diabetes? What is the number?**

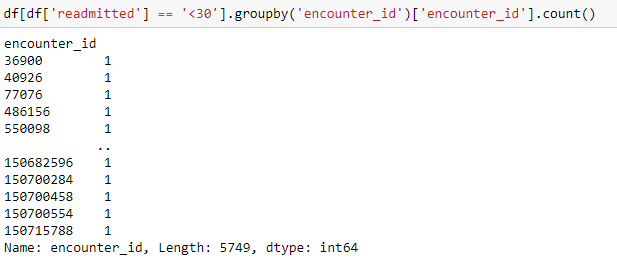
Answer: age between 70 and 80, 9817 people



**c. How many encounters whose admission type is Emergency? How many of these emergency encounters are readmitted within 30 days?**

Ans:24265 encounters, 5759 encounters

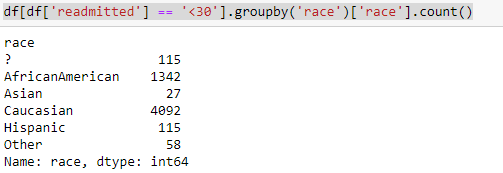


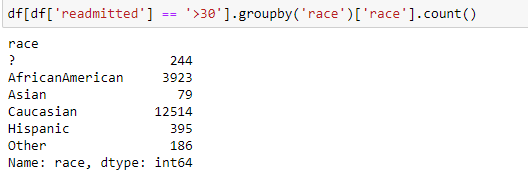


**d.What are the top-three race categories according to the number of readmission cases**

**(including both less than or larger than 30 days)?**

Answer for both less than or larger than 30 days are Caucasian, Africanamerican, Hispanic





**3. Using suitable visualisation plots:**

**1) Understand the distribution of variables and identify data quality problems.**

3.1.1 Weight variable has еthe maximum number of missing data is 48169 of a set of 49970 strings. It is unuseful to take this data into account for research.

3.1.2 For the gender column only one missing value is not a problem. We should analyze this data

3.1.3 The discharge\_disposition\_id variable has four missing data for id:21, 26, 29, 30 from 30 different. It makes sense to research this data.

3.1.4 The admission\_source\_id variable has 12 missing values for id:1, 12, 13, 15, 18, 19, 21, 22, 23, 24, 25, 26 ?

3.1.5 The payer\_code variable has еthe maximum number of missing data is 32665. In most cases this information is empty. It is unuseful to take this data into account for research.

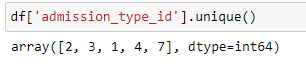
3.1.6 The medical\_specialty variable has еthe maximum number of missing data is 17755. It means one third of data is missing. It is unuseful to take this data into account for research.

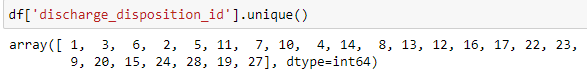
Variables 16, 17, 18 are only missing a small number of data. Hence, it doesn’t affect the whole data too much.

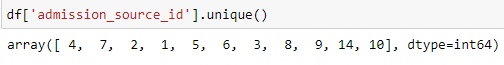
Variable 26 to 35 Over half half of data are “No”, which means too much data is not recorded. Thus these variables can't be used for further research.

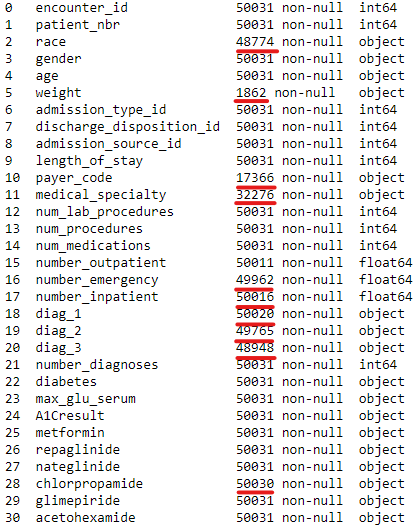
Variable 38 ‘readmitted’, over half data are not recorded, this variable can’t not be used for further research.

Variable 7,8,9 some of the number represent Not Available, NULL, Not Mapped, Unknown/Invalid



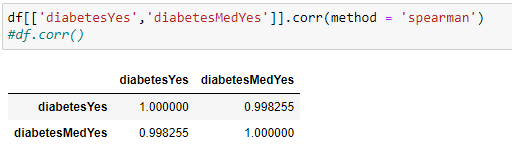




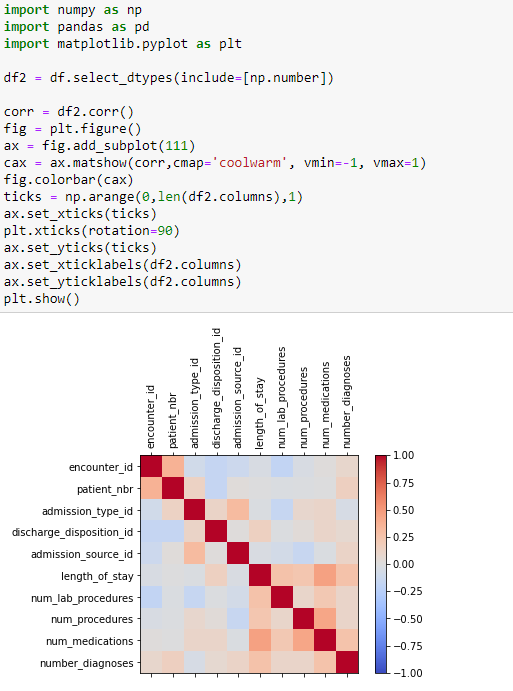
****

**2) Determine if there is any relationship between the variables diabetes and diabetesMed? How would you handle these two variables in the data modelling if a relationship exists?**

The relationship exists in these two variables, it is a highly positive relationship. These two variables have the same value. while doing the research or other works, these two variables should not be used at same time due to the weights of value will double.



**3) Identify the highly correlated variable pairs and elaborate on how to treat these variables in the mining process in such a case.**



The first pair is num\_procedures & num\_medications, and the second pair is num\_medications & Length\_of\_stay. These two pairs of variables are highly correlated in this dataset. For further data prediction, training, and testing, it is not suitable to use these at same time. Due to the weight of the value doubling, it might affect the result.

**4. Data preparation:**

**1) Summarize your findings based on data exploration.**

Variables with errors:

race variable has missing data for 1257 entries

gender variable has unknown data only for 1 entry

weight variable has the most data is unknown 48169 entries

discharge\_disposition\_id has no values for id=21,26,29,30

admission\_source\_id missing data has no values for id=1, 12,13,15,18,19,21,22,23,24,25,26

payer\_code has missing data for 32665 entries

medical\_specialty has missing data for 17755 entries

number\_outpatient has missing 20 entries

number\_emergency has missing 69 entries

number\_inpatient has missing 15 entries

chlorpropamide missing 1 data

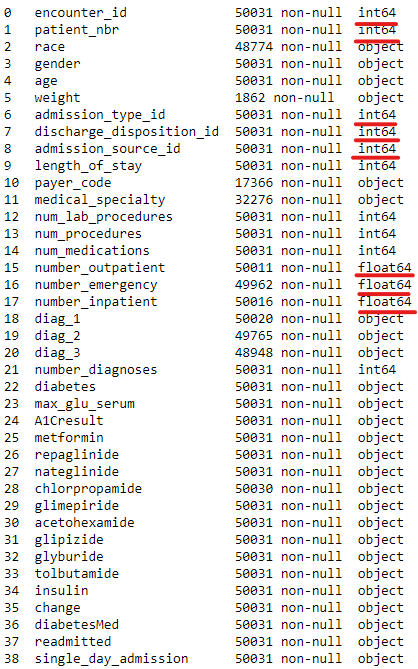
Variables with highly positive relationship:

1. diabetes and diabetesMed

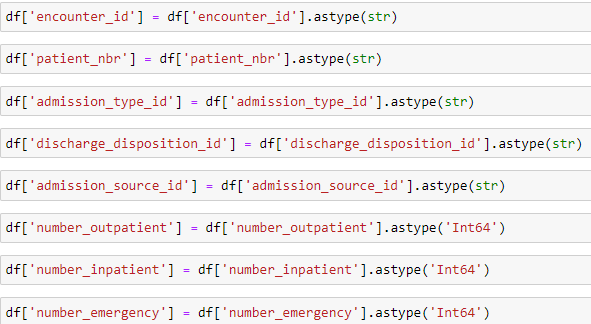
Highly correlated variables:

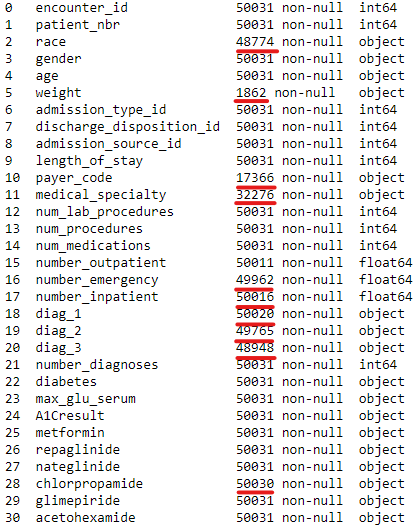
1. num\_procedures and num\_medications
2. num\_medications and length\_of\_stay

**2) Elaborate on the data preparation steps required (e.g., data cleaning and transformation) to address the data quality problems that you encountered during data exploration.**



For variables “encounter\_id”, “patient\_nbr”, “admission\_type\_id”, “discharge\_disposition\_id”, “admission\_source\_id”, the data type of these variables should be Nominal. “number\_outpatient”, “number\_emergency”, “number\_inpatient” should be Discrete.



****

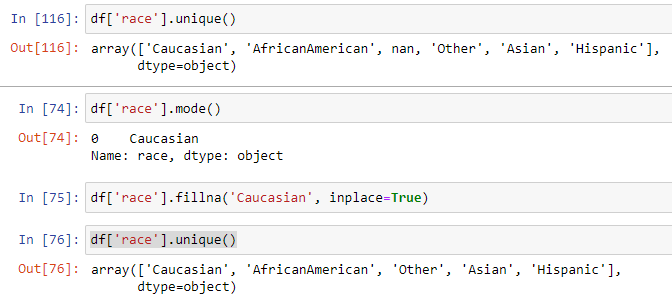
After correcting the data type, “.infor()” can indicate the information of the whole dataset, and it is clear to see that here are 11 variables that do not contain all values.

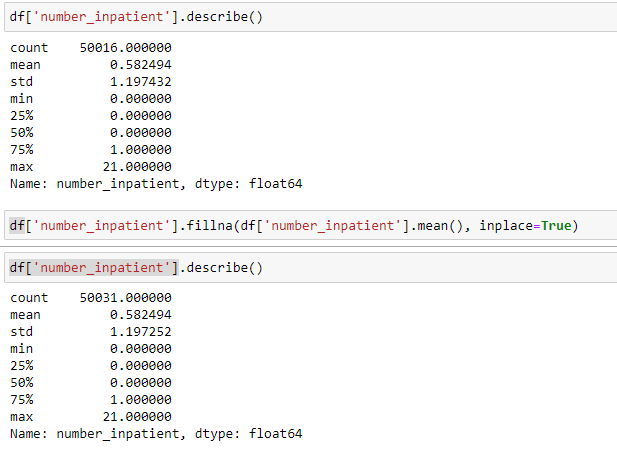
As for missing data, we use three different ways to deal with it: imputation, remove the whole column, drop the missing row.

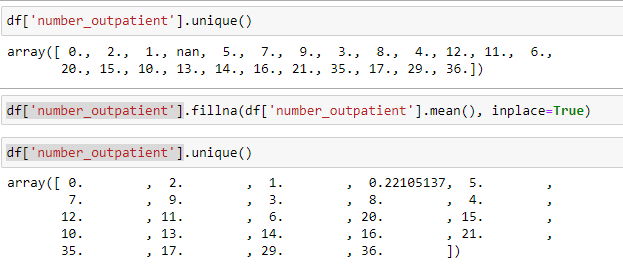
Imputation: due to the missing data in the column is only a very small number of it, we decided to fill mode value for Nominal variable, and mode value for Discrete variable.

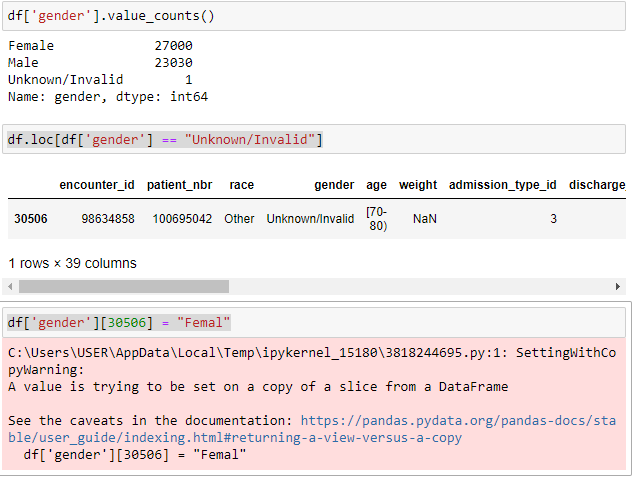
Remove column: the value of these three variables are representing different diseases. If we fill these values with any kind of value will affect the result.

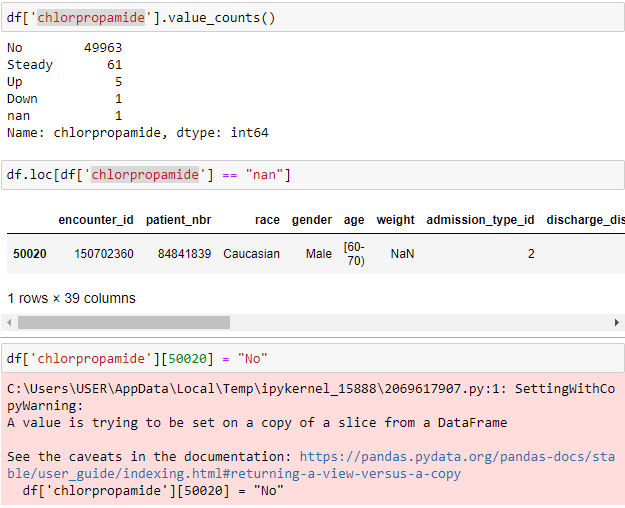
| imputation | Remove column |
| --- | --- |
| race  number\_inpatient number\_emergency  number\_outpatient  gender  chlorpropamide | weight  payer\_code  medical\_specialty  diag\_1  diag\_2  diag\_3 |

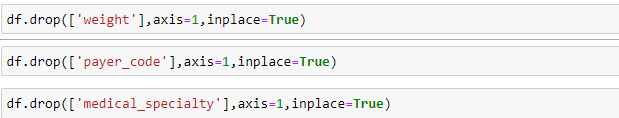


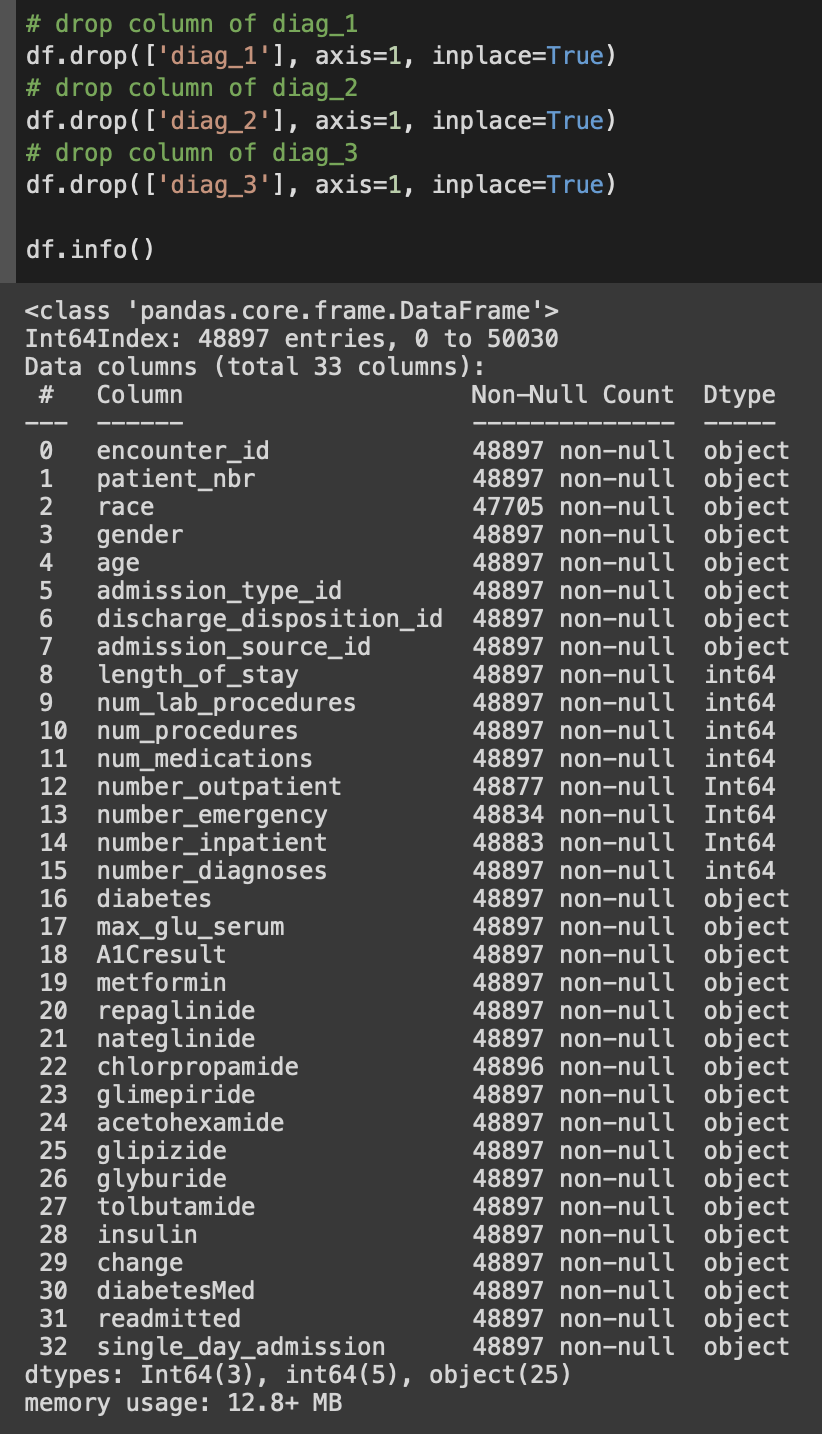




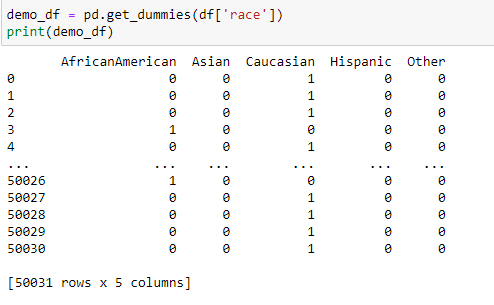


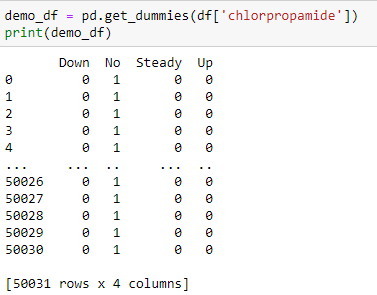


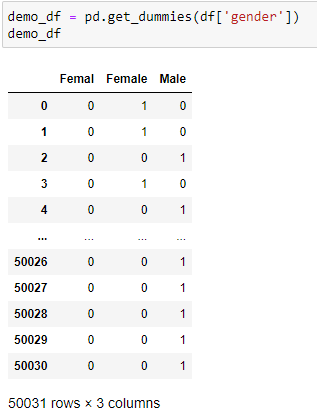


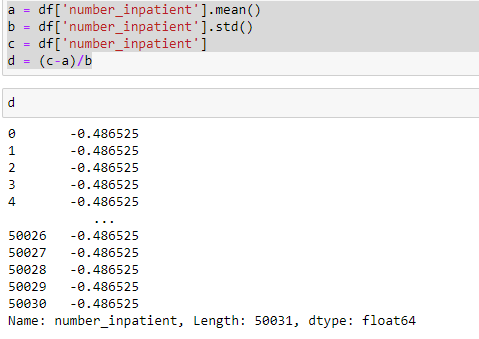


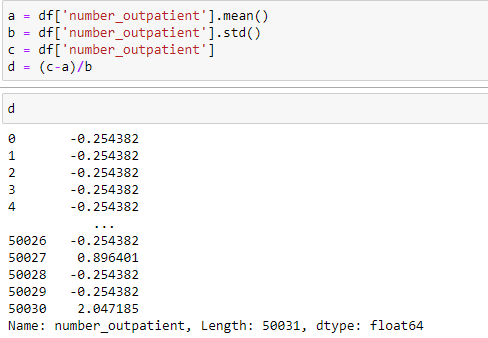
Data Transformation

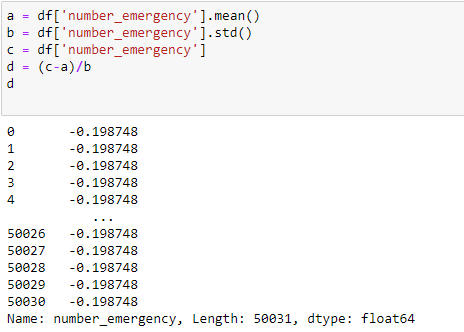




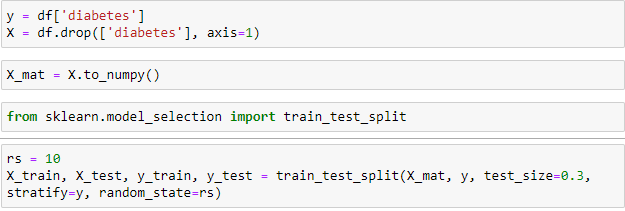








**3) Demonstrate the data preparation by including a screenshot(s) of the Python code and its outputs that show the steps on how you corrected all the identified data quality problems in this dataset.**



**Task 5**

**1) Identify the most suitable data mining task (i.e. classification, clustering or association mining) that can be performed on this dataset. Justify your choice.**

**2) What variables will you include in this data mining task and why? Describe here if you will create any derived variables. Identify the roles (i.e., input or target) of each variable.**

We think association mining is the most suitable data mining task of this dataset. For example, the ten variables (No.26-No.35) record the medication status. Hence, we can use these variables and variable “diabetes” (No.23) to figure out the relation between these variables. Moreover, the combination of these medications can show which will treat diabetes well.