IFN 509: Data Exploration and Mining

Assessment 1

Team Name: [ Team1088 ]

Group No. [ 17 ]

|  |  |
| --- | --- |
| Student Name | Student Id |
| Yeon Jae Park | n10886249 |
| Timothy Jeoung | n10887601 |
|  |  |

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

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

Question 1.

***Examine the data types assigned by the ‘Pandas’ library for each variable in Table 1. Check them with the dataset description provided in Table 1. If there is a mismatch in the data type assigned by the library and the data type as per the description for a variable, correct the data type. Report them. Attach a screenshot showing the correct data types of variables.***





The datatype was identified through ‘.info()’ function, which shows the datatypes of each variable. According to the above screenshot of the dataset and data description, there are mismatched data types which are “change” and “diabetesMed”.These two data types were assigned as objects, but since they only have two categories for each variable, they can be considered binary values. Therefore, the data type was changed to binary using a ‘map’ function that changed the ‘object’ data type to ‘int64’.

텍스트이(가) 표시된 사진

자동 생성된 설명

Table

Description automatically generatedText

Description automatically generated

Moreover, **diag\_1**, **diag\_2** and **diag\_3**should be the first three digits of the ICD9 code, meaning they should be integers, but they had object datatypes instead. Some of the ICD9 codes contain letters such as ‘V’ or ‘E’ and decimal numbers depending on their category. Data types of these variables were fixed later during data cleaning to be able to be used to calculate correlation with other variables.

Question 2.

***1) Identify and report the skewness present in the variables.***

The skewness of data can only be found within quantitative data, and hence the following variables were used to identify the skewness:

1. time\_in\_hospital
2. num\_lab\_procedures
3. num\_procedures
4. num\_medications
5. number\_outpatient
6. number\_emergency
7. number\_inpatient
8. number\_diagnoses

Mean and median of each data were used to find the skewness, since if

(mean – median) > 0, it is positively skewed and negatively skewed if less than 0.

Overall, the result was following:

텍스트이(가) 표시된 사진

자동 생성된 설명

|  |  |
| --- | --- |
| time\_in\_hospital | Postively skewed |
| num\_lab\_procedures | Negatively skewed |
| num\_procedures | Postively skewed |
| num\_medications | Postively skewed |
| number\_outpatient | Postively skewed |
| number\_emergency | Postively skewed |
| number\_inpatient | Postively skewed |
| number\_diagnoses | Negatively skewed |

***2) There may be inconsistencies or errors in the data. List the errors identified and***

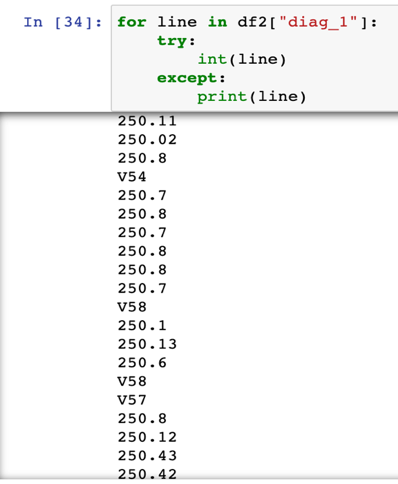
***detail how you have identified them.***

Several errors were identified using ‘.value\_counts()’ function for each qualitative variables. By using the function there were five variables with missing values, which were **race**, **weight**, **payer\_code**, **medical\_specialty**, **chlopropamide**, **diag\_1**, **diag\_2** and **diag\_3**. In addition, **diag\_1**, **diag\_2** and **diag\_3** also had inconsistencies within the data as mentioned before. All of the missing values and inconsistencies of qualitative variables were identified using .value\_counts() or .unique() function.

테이블이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명

테이블이(가) 표시된 사진

자동 생성된 설명

테이블이(가) 표시된 사진

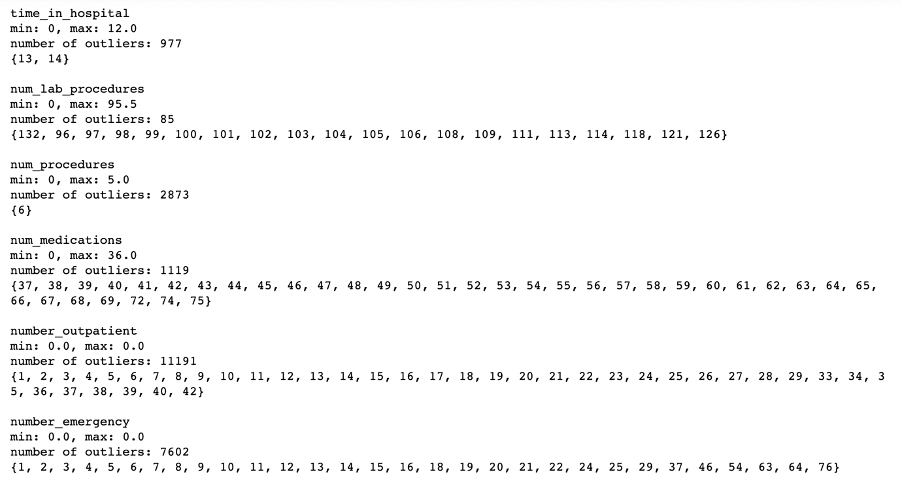
자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명

Furthermore, outliers and their count were found within quantitative variables, which were identified by calculating minimum value and maximum value for each variable using the equations below.

Q1 and Q3 were found using .describe() function, and if there were any values above maximum value and below minimum value, they were counted and added to set() to identify outliers. Screenshots below describe the outlier and the number of them.

텍스트이(가) 표시된 사진

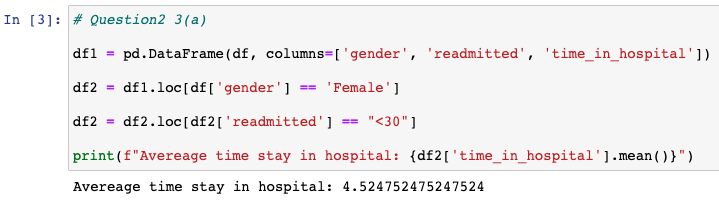
자동 생성된 설명

***3) Answer the followings:***

**A) What is the average time stay in hospital of a female patient who was**

**readmitted in less than 30 days?**

To solve this question, a new data frame was made with **gender**, **readmitted** and **time\_in\_hospital**. Then from the new dataset, females who were readmitted within 30 days were located using .loc and the mean of **time\_in\_hospital** was found. The average time of stays in hospital is approximately 4.52 days (Below is the approach to the answer).



**B) Which age group has the highest risk of being readmitted within 30 days?**

텍스트이(가) 표시된 사진

자동 생성된 설명

A new dataset was created with **age**and **readmitted,** and missing values were filled with the mode of the variable. Data with “<30” for “readmitted” were located and the count of each category was displayed in descending order using .value\_counts() and .sort\_values().

The age group with the highest risk of being readmitted within 30 days is [70-80].

**C) How many age groups have more than 3000 cases of being readmitted?**

From the dataset used in the previous question, patients who have been readmitted (**readmitted** that are not “NO”) were located, and the sum of each age group was found by using .groupby() and .sum(). Then the age groups were sorted based on their count.

Four age groups had more than 3000 cases of being readmitted, which were [70-80], [60-70], [80-90] and [50-60].

테이블이(가) 표시된 사진

자동 생성된 설명

**D) Which are the top-three race categories according to the number of readmission cases?**

Like previous questions, after creating a new dataset with **race** and **readmitted**, patients who were readmitted were located and grouped by their race to count the total number. Top three races were displayed using .sort\_values(ascending=False) and .head(3). The top three race categories that had the highest number of readmission cases were “Caucasian”, “AfricanAmerican” and “Other”.

텍스트이(가) 표시된 사진

자동 생성된 설명

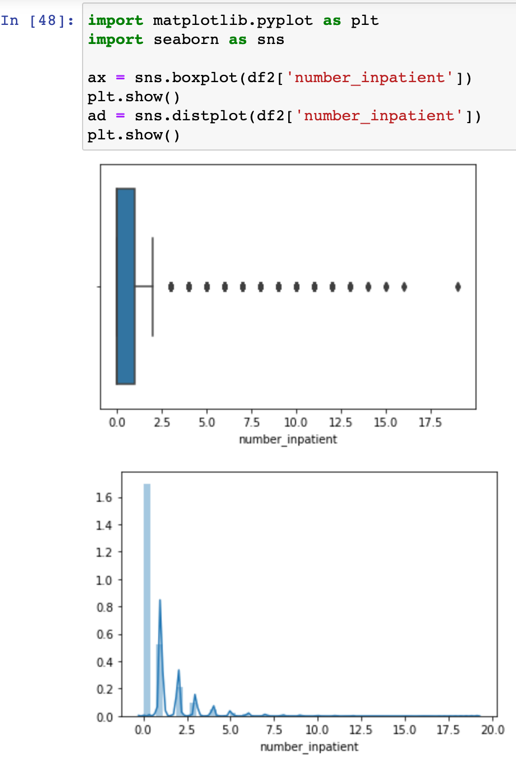
Question 3.

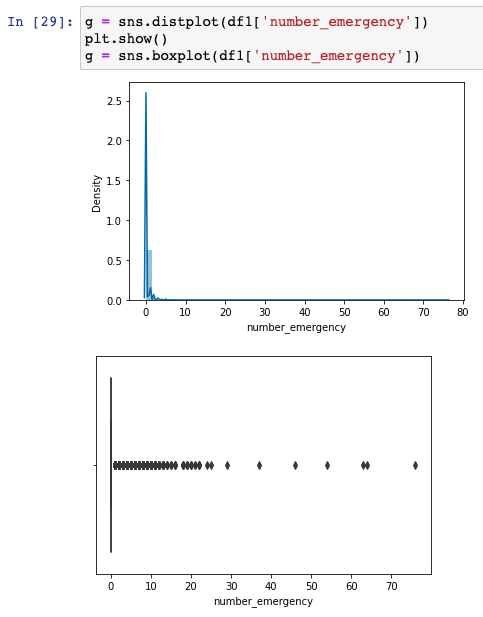
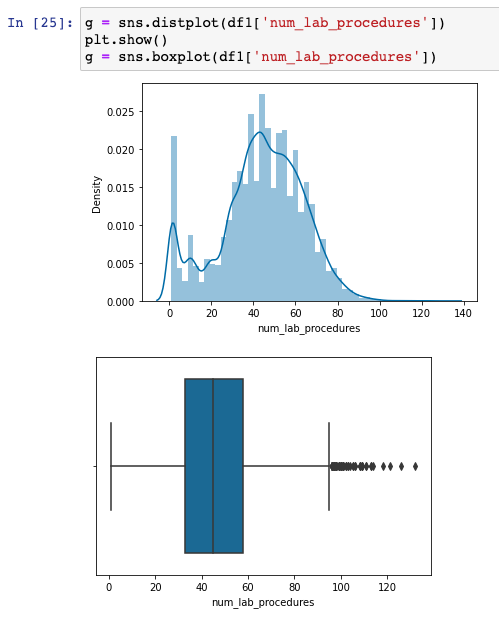
***1) Understand the distribution of variables and identify data quality problems.***

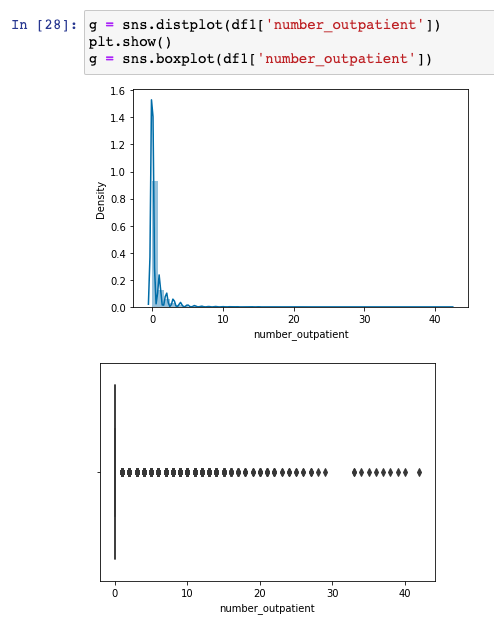
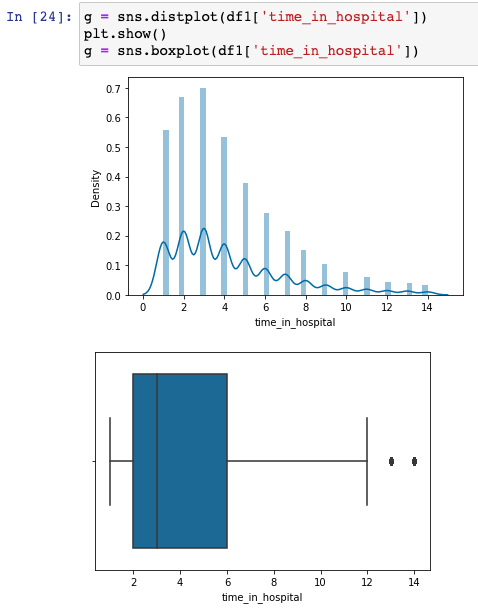
To understand the distribution of variables clearly and identify data quality problems, it is ideal to use visualisation techniques such as ‘boxplot’ and ‘distplot’. Data with interval variables were plotted for further investigation.

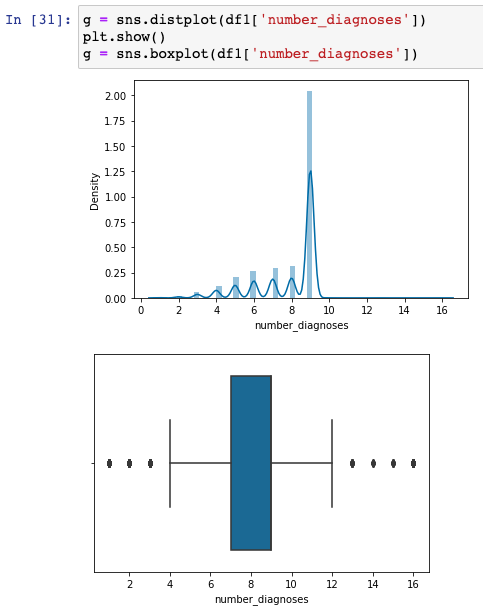
**time\_in\_hopital**, **num\_procedures** and **number\_diagnoses** showed a reasonable number of outliers and distribution, but other variables such as **num\_lab\_procedures**, **number\_outpatient, number\_inpatient** and **number\_emergency** showed a high number of outliers because most of the values were distributed in 0, making mean and median lower, creating more outliers. Ouliter could indicate patients with serious conditions who needed more prescriptions, tests or visits than other patients.

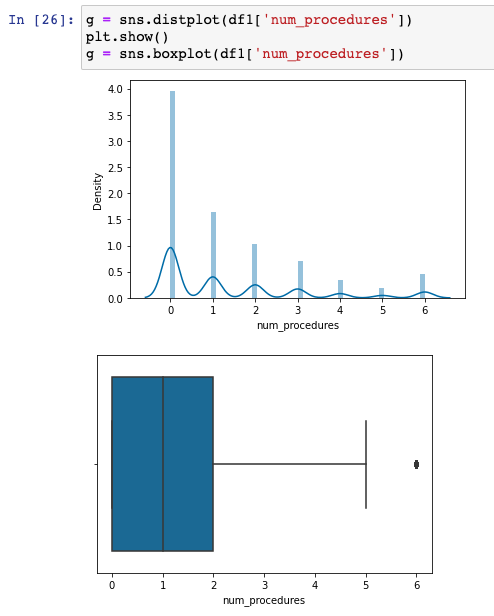
Chart, histogram

Description automatically generated





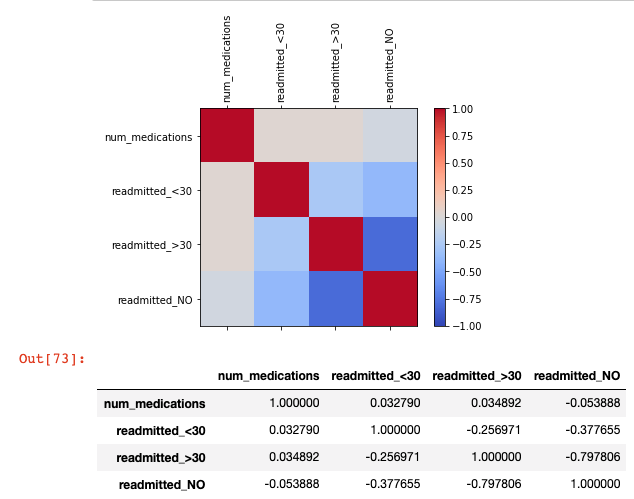




***2) Determine if there is any relationship between the variables num\_medications***  ***and readmitted within 30 days? How would you handle these two variables in*** ***the data modelling if a relationship exists?***

텍스트이(가) 표시된 사진

자동 생성된 설명

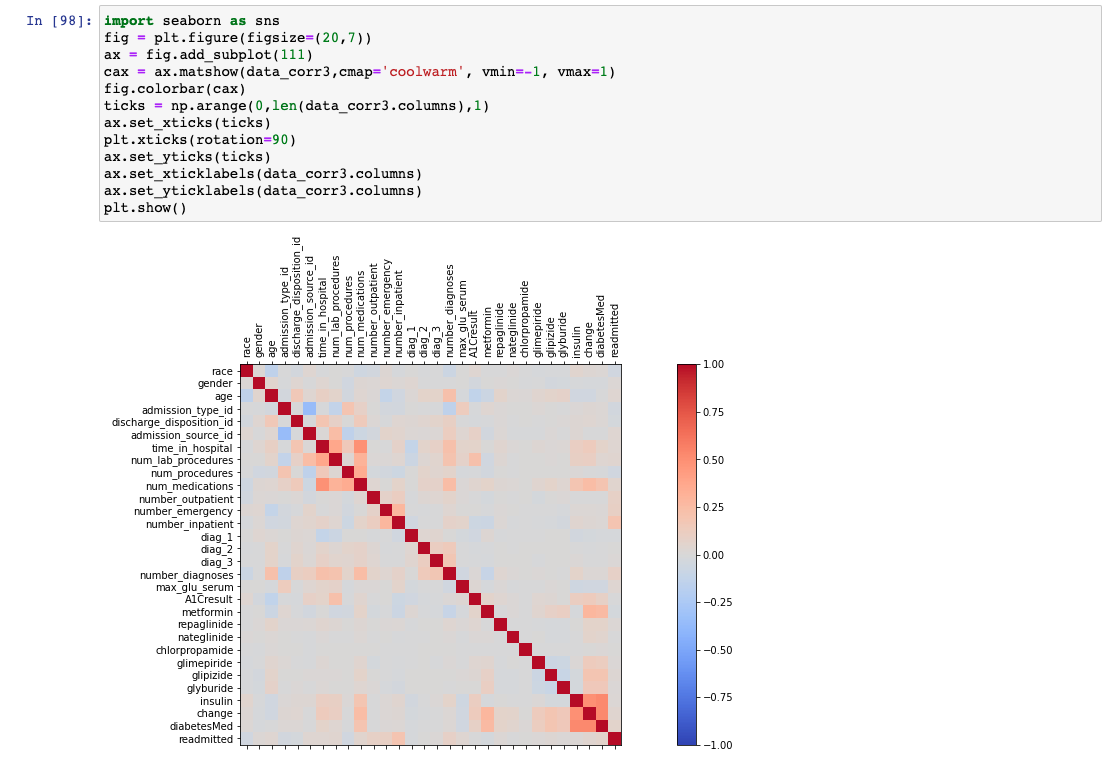


Using one hot encoding, values of **readmitted** were separated into three different binary values. The correlation between **num\_medications** and **readmitted\_<30** was 0.032790.

Since the correlation is very close to 0, it can be said to have a very small correlation (or none) between these two variables. If there was a relationship existing between the two variables, **num\_medications** can be used as an input variable and ‘readmitted’ as the target variable in data modelling.

***3) Identify the highly correlated variable pairs and elaborate on how to treat these***

***variables in the mining process in such a case.***



The correlation of the data frame was plotted to identify highly correlated variable pairs, but as shown in the graph, only a small number of variable pairs showed a positive correlation. For an accurate investigation, variable pairs that have relatively high correlation was printed out using for loop. Since there were not many values over 0.5, a correlation above 0.4 was displayed as following:텍스트이(가) 표시된 사진

자동 생성된 설명

As shown in the result above, there were four pairs of variables that have relatively high correlation compared to other pairs in the data frame, which were **time\_in\_hospital** with **num\_medications** (approx. 0.49), **insulin** with **change** (approx. 0.5), “insulin” with **diabetesMed** (approx. 0.53) and **diabetesMed** with **change** (approx. 0.53).

**Time\_in\_hospital** and **num\_medications** can be used for data modelling for prediction because high correlation would mean that one of them is a target variable, while another being an input variable. If the correlation was low between two variables, it would indicate that they are both input variables as they are independent.

On the other hand, the relationship between **insulin**, **diabetesMed** and **change** has a possibility of being a false predictor. Although they are highly correlated to each other, they describe events that are happening at the same time. They indicate a change in diabetes medications or prescription of them. For example, if there was an increased dosage of insulin, it would also indicate a change in dosage of diabetics medications, as well as meaning that the medication was prescribed. Therefore, they are not suitable to be used as predictors because other variables may be ignored by the model.

The main focus of the data modelling, in this case, is to predict the risk of the patient. Therefore, it would be ideal to select **time\_in\_hospital** as the target variable and then select other input variables for modelling.

Question 4.

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

Based on data exploration, there were some interesting findings identified within the dataset. First of all, variables such as **num\_medications, num\_lab\_procedures** and **num\_diagnoses** showed a relatively high correlation with **time\_in\_hospital** compared to other variables. A high correlation was found between **insulin, change** and **diabetesMed,** but thesehad events that were happening at the same time, and these can be considered false predictors and redundant variables, which should be removed during data reduction. Some of the variables had missing values that could be imputed with other values for better modelling of the data. However, variables such as **weight** and **medical\_specialty** were missing the majority of values and hence it would be ideal to exclude these variables for modelling. Finally, some of the interval variables showed a high number of outliers because most of the values were distributed in 0, which were smoothed for greater performance.

***2) Elaborate on the data preparation steps required (by correcting the data types of some variables, data cleaning and transformation) to address the data quality problems that you encountered during data exploration.***

**Data Selection**

In the data selection stage, unique variables, variables with the majority of missing values and irrelevant variables were dropped such as **encounter\_id, patient\_nbr, payer\_code, medical specialty** and **weight**.

**Data Cleaning: Improving Quality**

In the data cleaning stage, variables with incorrect data types were converted to proper data types. Binary variables (**change** and **diabetesMed**) that had a data type of object were converted to integers, and **diag\_1, diag\_2** and **diag\_3** had object data type due to inconsistent data. Hence values that had letters and decimals were fixed to have a whole number and then converted to an integer.

The next step was to identify missing values for each variable and clean them by filling the mode of the variable into missing values. Six variables (**race, gender, chlorpropamide, diag\_1, diag\_2 and diag\_3)** had missing values indicated as “?”, hence .replace() function was used instead of .fillna().

Finally, smoothing outliers for interval variables were done using z-score normalisation. For each normalised interval variable, .where() function was used to replace values that had absolute values more than 3 to mean of z-score. Therefore, all the outliers could be removed as well as normalised to a specific range.

**Data Transformation**

Variables with object data types were converted to integers, which can lead to better predictive performance and become more efficient. For example, in Question 1, **change** and **diabetesMed**  were changed to binary values from object data type. All of the categorical variables were converted to integers using .map() function, assigning a suitable number for each category. Overall, all the variables in the dataset ended up with numerical data types, which will allow calculating correlation for modelling.

**Data reduction**

After calculating correlation for the dataset using .corr(), some variables were not suitable as predictors since they only consist of one value or they were redundant. **Acetohexamide** and **tolbutamide** were removed as they only had “NO” as their values. **Insulin**, **diabetesMed** and **change** were highly correlated to each other, but they all described diabetic medication prescription or dosage. As **insulin** was a specific type of diabetic medication, it was kept in the dataset, whereas **change** and **diabetesMed** were dropped from the data. Overall, there were 28 variables remaining in the final pre-processed dataset to be used for modelling.

***3)*** ***Demonstrate the data preparation by including a screenshot(s) of the Python code***

***and its outputs that show the steps on how you had corrected all the identified data***

***quality problems in this dataset.***

1. **테이블이(가) 표시된 사진

   자동 생성된 설명텍스트이(가) 표시된 사진

   자동 생성된 설명Dropping variables & change data types**

As shown in Question 2.2, there were inconsistent and missing values for **diag\_1**, **diag\_2** and **diag\_3**, which were cleaned to be digits using a function to suit the description of variables.

텍스트이(가) 표시된 사진

자동 생성된 설명

Output:



1. **Filling missing values**

텍스트이(가) 표시된 사진

자동 생성된 설명

1. **테이블이(가) 표시된 사진

   자동 생성된 설명Smoothing interval variables by using z-score normalisation**
2. **Data Transformation**

Variables with data types of object were converted to integers to be used for modelling and calculating correlation. Mapping function was used for conversion as following:

텍스트이(가) 표시된 사진

자동 생성된 설명

테이블이(가) 표시된 사진

자동 생성된 설명

1. **Data Reduction**

**Acetohexamide** and **tolbutamide** were not suitable as predictors since they only consist of one category, which can be dropped from the dataset. Besides, **insulin**, **diabetesMed** and **change** are redundant, hence **diabetesMed** and **change** were dropped as well. Overall, there were 28 variables left in the final pre-processed dataset to be used for data modelling.

Question 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.***

Data mining can be categorised into two types, descriptive data mining, which is used to find human-interpretable patterns or clusters to describe the data, and predictive data mining for predicting future behaviour based on other variables. Classification is a predictive data mining technique that can predict large datasets accurately, whereas clustering and association mining are descriptive. The main goal of this dataset is to predict the risk of patients. Therefore, this dataset's most suitable data mining task would be classification rather than clustering and association mining.

***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.***

Using classification for data mining task requires categorical variables as the target variable. **Time\_in\_hospital** was selected as the target variable after pre-processing the dataset, but since it is an integer variable, it should be converted to a categorical variable to be used as a target variable in classification. Variable **time\_in\_hospital**can be derived into a new variable by categorising into three variables, low risk, moderate risk and high risk of being in hospital for an extended period, using the binning method.

Several variables such as **num\_medications, num\_lab\_procedures**and **number\_diagnoses**, which have are relatively correlated to **time\_in\_hospital**compared to other variables. These variables can be used as input variables to be used directly as predictors and distinguish segments. Other variables that were kept after pre-processing the dataset can be used as supplementary variables to explain the results.