Exploratory Data Analysis

The data set comes from 130 US hospital and integrated delivery networks during 10 years between 1999 and 2008 in order to analyse and care the diabetic patient. This data set represents 55 features and 100,000 instances about patient and hospital outcomes.

The race, gender and age are patients’ information. Type of admission describes a numerical value between 1 to 8 to separate admission type by hospital criteria. Type of discharge also describes a numerical value between 1 to 28 to classify discharge type. The source of admission describes a natural number from 1 to 25 in order to identify the source because the hospital is different. Max\_glu\_serum, A1Cresult, metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, cutoglipton, insulin, glyburide\_metformin, glipizide\_metformin, glimepiride\_pioglitazone, metformin\_rosiglitazone, andmetformin\_pioglitazone are type of hormone so these describe four types of status such as none, down, steady and up. These status covert to numerical values such as 1, 2, 3 and 4. Time in hospital and the number of patients’ information describe the natural number because these data is to count number.

As seen from the below table, the variables are reflected statistical information such as mean, median, minimum, maximum and quartiles. The variables in raw data is composed of character’s data but the data is labelled by a numerical value for example, female is 1 and female is 2. After that, the applied data can be analysed. These can be described what data is. The number of variables is forty-two and converted numerical values are race, gender, age, max\_glu\_serum, A1Cresult, metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, cutoglipton, insulin, glyburide\_metformin, glipizide\_metformin, glimepiride\_pioglitazone, metformin\_rosiglitazone, metformin\_pioglitazone, change, diabetic medication and readmitted. These variables are able to identify what option is frequent and what data is represented.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Continuous variables | | | | | | | | | | |
|  | Type of discharge | Source of admission | Time in hospital | Number of lab test performed | | Number of medications | Number of outpatients | Number of emergencies | Number of inpatients | Number of diagnoses |
| min | 1 | 1 | 1 | 1 | | 1 | 0 | 0 | 0 | 3 |
| max | 28 | 25 | 14 | 132 | | 81 | 42 | 76 | 21 | 16 |
| median | 1 | 7 | 4 | 44 | | 15 | 0 | 0 | 0 | 8 |
| Mean | 3.75 | 5.78 | 4.42 | 43.15 | | 16.120 | 0.376 | 0.202 | 0.647 | 7.512 |
| std.dev | 5.309 | 4.071 | 2.993 | 1.971 | | 8.108 | 1.283 | 0.943 | 1.271 | 1.832 |
| Categorical variables | | | | | | | | | | |
| Freq | Type of admission | Number of procedures |  | | | | | | | |
| 0 | - | 44574 |
| 1 | 52178 | 20029 |
| 2 | 17543 | 12383 |
| 3 | 18194 | 9210 |
| 4 | 10 | 4076 |
| 5 | 4661 | 2970 |
| 6 | 5135 | 4811 |
| 7 | 20 | - |
| 8 | 312 | - |
|  | | | | | | | | | | |
| freq | max\_glu\_serum | A1Cresult | Metformin | Repaglinide | | Chlorpropamide | Glimepiride | Acetohexamide | Glipizide | Glyburide |
| 1 | 92845 | 81860 | 78808 | 96530 | | 97970 | 93066 | 98052 | 85769 | 87792 |
| 2 | 2532 | 4854 | 551 | 45 | | 1 | 184 | 0 | 541 | 538 |
| 3 | 1449 | 3708 | 17677 | 1371 | | 76 | 4488 | 1 | 10991 | 8932 |
| 4 | 1227 | 7631 | 1017 | 107 | | 6 | 315 | 0 | 752 | 791 |
| 5 | - | - | - | - | | - | - | - | - | - |
|  | | | | | | | | | | |
| freq | tolbutamide | Pioglitazone | Rosiglitazone | Acarbose | | Miglitol | Troglitazone | Tolazamide | Examide | Citoglipton |
| 1 | 98031 | 90955 | 91887 | 97754 | | 98016 | 98050 | 98016 | 98053 | 98053 |
| 2 | 0 | 115 | 84 | 3 | | 4 | 0 | 0 | 0 | 0 |
| 3 | 22 | 6756 | 5908 | 286 | | 31 | 3 | 36 | 0 | 0 |
| 4 | 0 | 227 | 174 | 10 | | 2 | 0 | 1 | 0 | 0 |
| 5 |  | - | - | - | | - | - | - | - | - |
|  | | | | | | | | | | |
| freq | Insulin | Glyburide.metformin | Glipizide.metformin | Glimepiride.pioglitazone | | Metformin.rosiglitazone | Metformin.pioglitazone | Change | Diabetic medication | Readmitted |
| 1 | 45943 | 97384 | 98040 | 98052 | | 98053 | 98052 | 52774 | 22702 | 52338 |
| 2 | 11843 | 3 | 0 | 0 | | 0 | 0 | 45279 | 75351 | 11066 |
| 3 | 29368 | 660 | 13 | 1 | | 0 | 1 | 0 | 0 | 34649 |
| 4 | 10899 | 6 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 |
| 5 | - | - | - | - | | - | - | - | - | - |
|  | | | | | | | | | | |
| freq | race | Gender | nateglinide | |  | | | | | |
| 1 | 75079 | 52833 | 97362 | |
| 2 | 18881 | 45219 | 11 | |
| 3 | 625 | 1 | 657 | |
| 4 | 1984 | - | 23 | |
| 5 | 1484 | - | - | |
|  | | | | | | | | | | |
| freq | age |  | | | | | | | | |
| 1 | 65 |
| 10 | 466 |
| 20 | 1478 |
| 30 | 3548 |
| 40 | 9265 |
| 50 | 16697 |
| 60 | 21809 |
| 70 | 25306 |
| 80 | 16702 |
| 90 | 2717 |
|  | | | | | | | | | | |

*Table 1: Separated two part, continuous and categorical variables to describe statistics.*

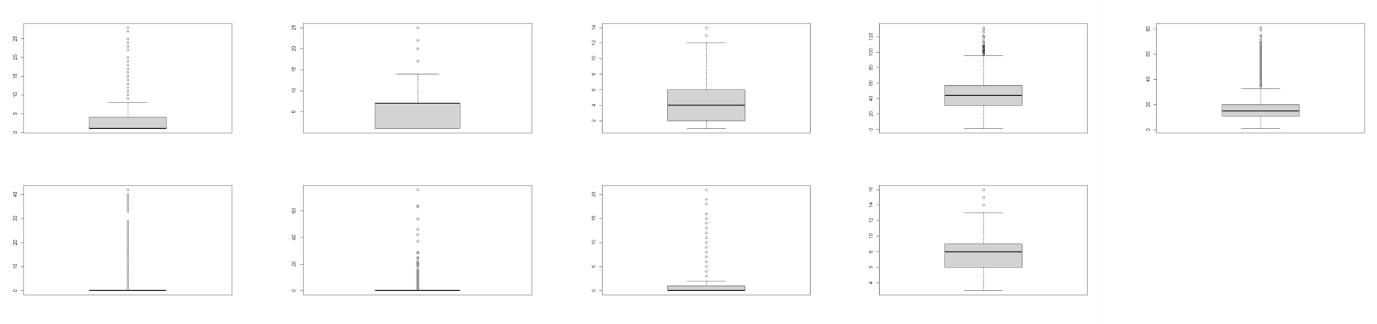
The variables are categorized such as categorical and continuous values. The number of rows about each variable is 98,053 which is removed missing values from raw data. Firstly, the race is a categorical variable. The most frequency of categories is 75,079 as 1 and the percentage of Caucasian is 76.57% so the Caucasian can be significantly taken up the data set. Next, the ratio of females can be higher than males because the most frequency of female is 52833, it is closed to 1 as labeling female and the percentage of female is 53.88%. The age is located between 0 and 100. The 70s are often distributed, 25306 people age is 70s. Therefore, the 50s, 60s, 70s and 80s ages can be frequently indicated more than others. The max\_glu\_serum column’s most frequency is 92845 as 1 and the percentage of 1 is 94.69%. This means that 1 as ‘none’ makes up most of the column. Some frequencies of columns are mostly to 1 that means these columns can have so many ‘none’ values. None of A1CResult can be frequently indicated for the above table because A1CResult’s most frequency value is 81860 and the percentage of 1 is 83.49%.

The following categorical variables are how to maintain to take treatment such as no, down, steady and up. 1, 2, 3 and 4 are represented the number of categorizations of each column. First, an frequency of metformin is 78808 as 1 and the percentage of no is 80.37%. It means that patients cannot take medicine such as metformin. When patients take repaglinide, the repaglinide cannot be taken because the most frequency is 96530 as 1 and the percentage of no is 98.45%. In addition to the chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, cutoglipton, glyburide\_metaformin, glipizide\_metaformin, glimepiride\_glitazone, metformin\_rosiglitzone, metformin\_pioglitazone, these’ frequencies are mostly frequencies as 1 and the percentages of these are more than 90%. It means that patients cannot take treatment about the above list. However, the percentage of insulin column cannot be more than 50%. The most frequency is 45953 as 1 and the percentage of insulin is 46.87% as well as the other categories have more than 10,000 thus the data can be differently distributed.

Moreover, the rests of categorical variables are change, diabetic medication and readmitted column. The change and diabetic medication columns are binary option. The most frequency of change column is 52774 as 1 so some patients change. Furthermore, the diabetic medication’s most frequency is 75351 and the percentage of diabetic medication is 77.03%. It means that the values are often represented 2 which is labelled yes. The most data is located at 2. Specifically, readmission is focused on the frequency is 52338. This value is not often 1 unlike other categorical variables which 2 is labelled by more than 30 days and 3 is less than 30 days. The patients will be possible to readmit more and less than 30 days. These data can be distributed for three options. As a result, the histogram and plot are necessary in order to deeply analyse.

On the other hand, the continuous variables are type of admission, type of discharge, source of admission, number of lab test performed, number of procedures, number of medications, number of outpatients, number of emergencies, number of inpatients, and number of diagnoses. First, the type of admission average is 2.03 and the median is 1. The values in type of admission column cannot spread out and 1 can be frequently indicated. Next, type of discharge column can have a lot of 1 because this column median is 1 and the mean is 3.75 even though the maximum value is 28. Like type of admission and type of discharge, source of admission data can be crowded to 5 because mean is 5.78 and median is 7.

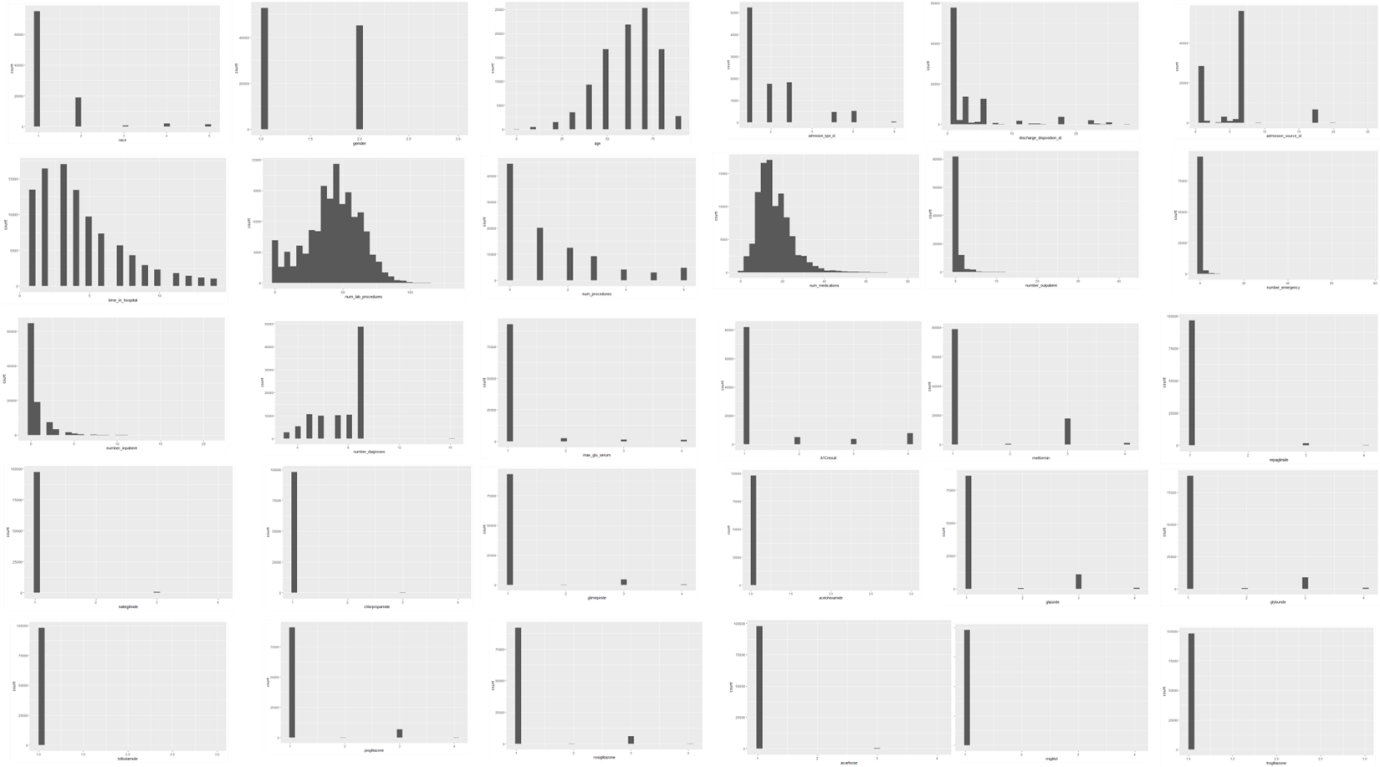
Furthermore, the rests of the continuous variable can identify the diabetic patients’ pattern. The time in hospital column average is 4.42 and the median is 4. It means that patients can go to the doctor about 4 times and some patients can visit 14 times. Overall, the diabetic patients proceed laboratory about 43 times by identifying median 44, mean 43.15. Some patients proceed highly for 132 times. However, the number of procedures is 1.35 on average thus patients are proceeded about a time in general as well as patients medicate for approximately 16 and maximum of medication is 81 and at least, some patients medicate for a time. The number of outpatients, emergency and inpatient can be less than 1 so patients cannot occur three categorizations. Finally, the patients are diagnosed for about 8 times in general because the mean is 7.512 and median is 8.

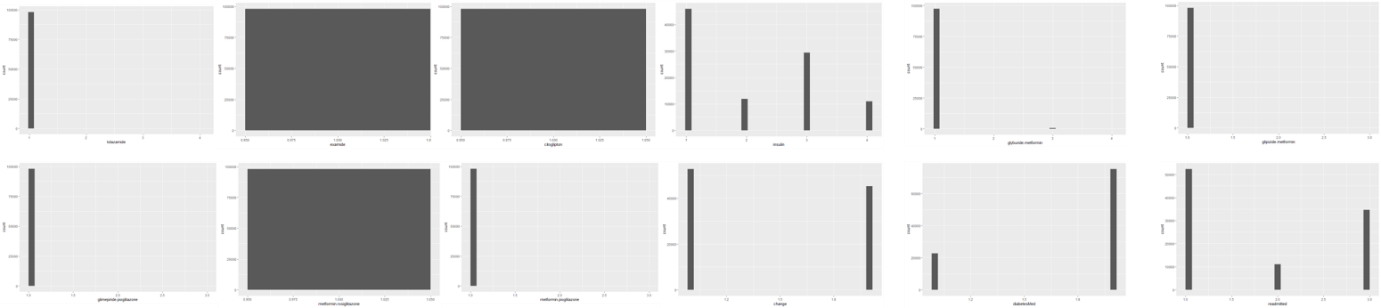


*Figure 1: boxplot of the continuous variables.*

As seen from the Figure 1, the continuous variables are depicted as boxplot. Boxplot can be identified data distribution, median, outlier, skewness, and dispersion. The box can be indicated interquartile range between first quartile and third quartile, and the bold line in box is median (2nd quartile). Moreover, it is possible to identify the data can be statistical distribution such as normal distribution as well as tiny circle is represented outlier.

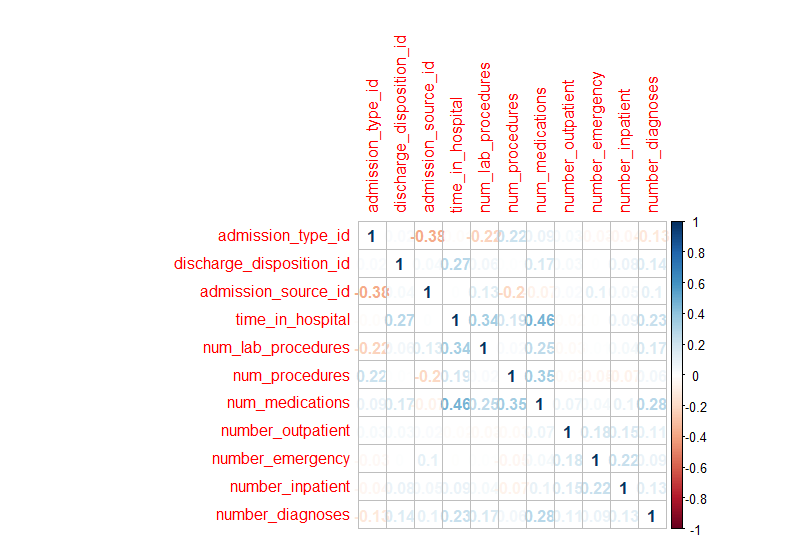
Time in hospital and number of medications can have little to no outliers. Number of lab test performed, number of medications, number of inpatients and number of diagnoses can be normal distribution, thus these four columns can be significantly distributed. Type of discharge and time in hospital can be right skewed. However, source of admission can be left skewed. Finally, box plots can be described variation. Type of discharge, number of medication, number of outpatients, number of emergencies and number of diagnoses columns have relatively a small variation, and others are comparatively larger. The larger variation’ columns can be spread out and distributed.





*Figure 2: Histogram of all of the variables, starting with 1 at top left.*

For the Figure2, each histogram represents how often the data is located. Specifically, categorical variables can be crowded for 1 so it can be matched above descriptive analysis table as the mean values can close to 1. However, continuous variables can be distributed and for the above second row, first second and forth column, these data set can be normal distribution. The histograms support above data analysis so the data in some column is crowded for one value and another can be identified which data is frequent and pattern.



*Figure 3: Correlation plot of continuous variables by using Spearman’s rank coefficient.*

As seen from the above plot, the correlation values are represented by using Spearman’s rank coefficient for continuous variables. Spearman’s rank coefficient can measure the rank correlation between the rankings of two variables like continuous variables. The data set cannot have a strong correlation but some variables can be related to each other such as time in hospital and the number of medications. Furthermore, the number of laboratory procedures and time in the hospital, and source of admission and type of admission, and the number of procedures and medication can be related each other despite weak relationship value between 0,3 and 0.5.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A1Cresult | glimepiride | acetohexamide | glyburide | pioglitazone | acarbose | miglitol | examide | change | Diabetic medication |
| A1Cresult | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0 |
| glimepiride |  | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0.1 |
| acetohexamide |  |  | 1 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0 |
| Glyburide |  |  |  | 1 | 0 | 0 | 0 | 0 | 0.2 | 0.2 |
| Pioglitazone |  |  |  |  | 1 | 0 | 0 | 0 | 0.1 | 0.1 |
| acarbose |  |  |  |  |  | 1 | 0 | 0 | 0.2 | 0.1 |
| miglitol |  |  |  |  |  |  | 1 | 0 | 0.1 | 0.1 |
| examide |  |  |  |  |  |  |  | 1 | 0.6 | 0.5 |
| change |  |  |  |  |  |  |  |  | 1 | 0.5 |
| Diabetic medication |  |  |  |  |  |  |  |  |  | 1 |

*Figure 4: Correlation plot of categorical variables by using Cramer V, selecting more over 0.1.*

For above Figure 4, this plot is represented relationships for categorical variables by using Cramer V test. Cramer V test can be based on Pearson’s chi-squared statistics, which measure the association between two nominal variables. These categorical variables cannot relate to each other because the value of the relationship can be indicated almost zero. Therefore, the columns are selected more over 0.1 such as A1Cresult, glyburide, pioglitazone, acarbose, miglitol, examide, change, diabetic medication. The change column has a little correlation with other 9 featuers. As a result, categorical variables are necessary to find a new pattern by developing analysed methods.