TRIAGE

Triage_02: Chong Yong Ming, Tan Tee Yin, Loh Kah Fung

I. Background Study and Aim

TRIAGE is the primarily risk assessment and a form of prioritization delegating task of patient care based on the illness/ injury, severity, prognosis, and resource availability. The purpose of triage is to identify patients that need an immediate attention and medical care. [1].

The primarily risk assessment focuses on the assessment of the patients' risk based on several metrics such as age, gender, chest pain, blood pressure, cholesterol, max heart rate, exercise angina, plasma glucose, skin thickness, insulin, BMI, diabetes pedigree, hypertension, heart disease, residence type, smoking status and categorizes them into level of triage whereby blue is denoting as least severity and does not require an immediate sought of medical attention whereas red is denoting as the most severe that required immediate attention from physician. The ordinal level goes from least to most severe by tagging with blue, green, yellow, orange and red. Louis et. al conducted a Triage study to describe the formation of the Triage database and characterize the included patients, they have developed a Triage algorithm for stratifying patients according to the acuity level with predefined definitive criteria and patients' data using Chi-square, One-way ANOVA or Kruskal Wallis H test to compare the data across the triage categories. [2]

Logically says, patient considered to be at low risk for the evaluation if in young age, no chest pain, normal range of blood pressure, cholesterol level, heart rate, insulin level, BMI etc. The correlation of these metrics is in ambiguity and therefore, there is a need to further exploit, understand and describe the pre-set dataset to well define the triage categorization by exploratory data analysis (EDA).

II. Methodology

A. Dataset Description

The Triage dataset is a multivariate, and the associated task is a multiclass classification. It consists of 6961 of instances and 17 attributes. Table I depicts a total of 9 numerical attributes which subdivided into discrete and continuous and 8 categorical attributes which subdivided into nominal and ordinal.

B. Data Treatment

First, the triage.csv dataset is loaded into R for data treatment to check for missing values and any possible outlier. Based on the table shown in Figure 1, there is a null attribute that has an

TABLE. 1. Data description of triageData

Columns	Type	Value/Statistics	Percentage of Missing Value
Age	Discrete numerical	Range: 28 – 82 Mean: 57.45 Std: 11.91	0.00%
Gender	Nominal categorical	0 – male (3258) 1 – female (3703)	0.00%
Chest pain type	Ordinal categorical	0 – No Chest Pain (5822) 1 – Typical angina (60) 2 – Atypical angina (209) 3 – Non-anginal pain (275) 4 - Asymptomatic (595)	0.00%
Blood Pressure	Discrete numerical	Range: 60 - 165 Mean: 109.4 Std: 21.46	0.00%
Cholesterol	Discrete numerical	Range: 150 - 294 Mean: 184.9 Std: 31.91	0.00%
Max heart rate	Discrete numerical	Range: 138 - 202 Mean: 163.2 Std: 15.38	0.00%
Exercise Angina	Nominal categorical	0 – No (6531) 1 – Present (430)	0.00%
Plasma glucose	Continuous numerical	Range: 55.12 - 199 Mean: 98.43 Std: 28.59	0.00%
Skin_thickness	Discrete numerical	Range: 21 - 99 Mean: 57.33 Std: 22.90	0.00%
Insulin	Discrete numerical	Range: 81 - 171 Mean: 110.5 Std: 17.54	0.00%
BMI	Continuous numerical	Range: 81 - 171 Mean: 110.5 Std: 17.54	0.00%
Diabetes_pedigree	Continuous numerical	Range: 0.0780 – 2.420 Mean: 0.4674 Std: 0.1027	0.00%
Hypertension	Nominal categorical	0 – No (6463) 1 – Present (498)	0.00%
Heart_disease	Nominal categorical	0 – No (6686) 1 – Present (275)	0.00%
Residence_type	Nominal categorical	Rural (2512) Urban (4449)	0.00%
Smoking_status	Nominal categorical	Formerly smoked (2512) Never smoked (4449) Smokes (789)	22.18%
Triage	Ordinal categorical	Blue – minor injuries or complaints (422) Green – non-urgent (440) Orange – emergent (339) Red – Immediate evaluation by physician (129) Yellow – potentially unstable (5631)	0.00%

automatically assigned name of X in the first column. According to the original dataset, X is the attribute to show the number of data and its sequence, however, R programming does include the number of id for each row in the dataset. Therefore, the X attribute is suggested to be removed to avoid any duplication of dataset that might has a negative impact on the data analysis on the classification of triage.

To remove duplicated or unimportant column/attribute, the command is used as below:

```
> triageData = subset(triageData, select = -c(X))
```

Next, datatype of the attribute from numerical/character to categorical factor by using the command below:

```
> triageData$xyz <- as.factor(triageData$xyz)
```

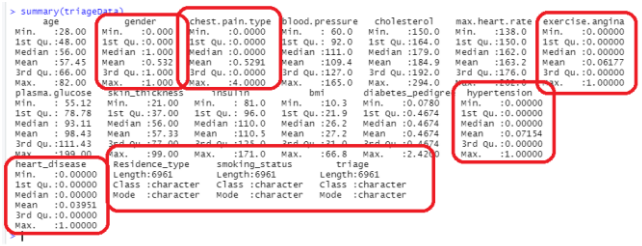

Fig. 1. Summary of the triage dataset attributes that require a datatype conversion

After completing all the datatype conversion for the attributes mentioned in Figure 1, the next step will be the checking process of missing values and outliers analysis.

For better data treatment and exploration, the value of "0" and "1" can be converted into female and male while the "unknown" status of smoking_status can be dropped by using the command below. Besides, based on the summary in Figure 2 shows there is no missing value in numerical data.

```
> triageData$gender = ifelse(triageData$gender==0, "Female", "Male")
> triageData$gender <- as.factor(triageData$gender)
> summary(triageData[ triageData$smoking_status != "Unknown", , drop=FALSE])
> triageData <- triageData[ triageData$smoking_status != "Unknown", , drop=FALSE]; triageData$smoking_status <- factor(triageData$smoking_status); summary(triageData)
```

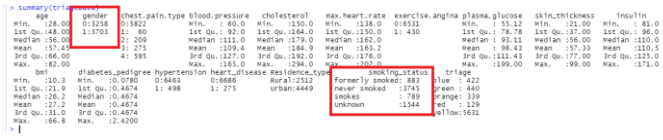

Fig. 2. Summary of missing values in categorical data.

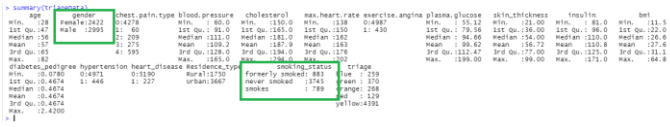

Fig. 3. Summary of triageData after data treatment.

C. *Data Exploration*

Triage is a classification of the emergency patients according to the urgencies such as Red (immediate evaluation by physician), Orange (emergent), Yellow (potentially unstable), Green (non-urgent), Blue (minor injuries or complaints). Based on Figure 3.5, we can notice that the distribution of the target attribute of triage is not evenly spread with yellow triage as most of the class with the number of 4391 while red triage is the least triage with the count of 129. This is not a good target dataset if we plan to use it for the machine learning model training and prediction. However, we can apply data augmentation to generate a more evenly distributed dataset among the target class.

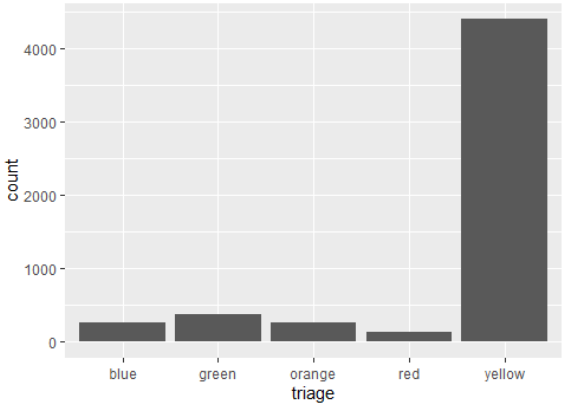

Fig. 4. Visualization of the distribution of target attribute triage with bar chart.

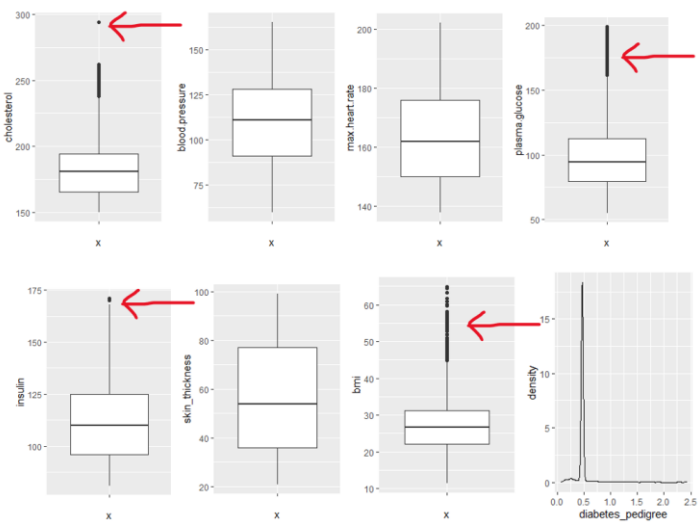

Fig. 5. Visualization of outliers with boxplot in R and density plot of diabetes_pedigree.

In addition, we can use visualization tool in R to help to determine the possible of outliers in the attributes of dataset using boxplot. Based on Figure 5, we can easily identify the outlier for cholesterol attribute, as well as with few of other attributes like BMI, insulin and plasmaglucose. Besides, we can also identify that there is an obvious density accumulated near around 0.5 for diabetes_pedigree with the density plot. Other than using visualization tools in R to identify outliers, formula or function can be used to calculate the range of outliers' values. The formula for lower range limit = Q1 – (1.5* IQR) and higher range limit = Q3 + (1.5*IQR). Therefore, any of the value located below or above these limits will be considered as an outlier.

For example, plasma.glucose. First, we need to calculate as below:

IQR=Q3-Q1,

IQR=112.47-79.56 = 32.91

Lower range limit = 79.56 – (1.5*32.91) = **30.91**

Higher range limit = 112.47 + (1.5*32.91) = **161.84**.

Refer to the max value of plasma.glucose in Figure 3, 199 is more than 161.84, which can be considered as outliers.

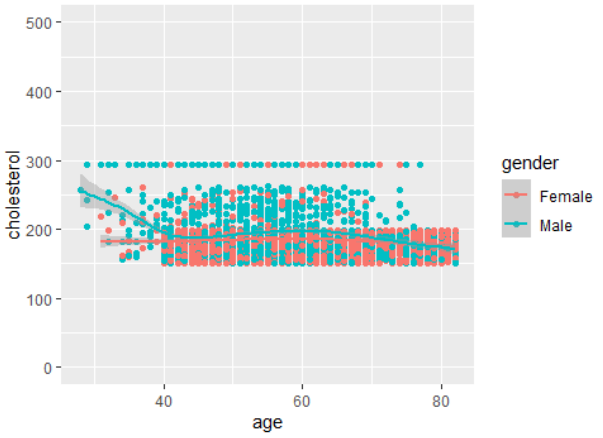

Fig. 6. Visualization of cholesterol against age that separated by gender.

Based on Figure 6, we can estimate that majority of the male are having higher cholesterol level as compared to the female. Besides, the there is no significant correlation between cholesterol and age. These insights are very useful for classifying the triage if the emergency cases are highly impacted by the cholesterol level.

III. Conclusion

In this article, the dataset described the triage based on various features which can be collected from the patient. The dataset consists of 17 attributes, 9 numerical and 8 categorical. To ensure the correctness and integrity of the dataset, data treatment is very important to handle the missing value, invalid value, and outliers. With the help of R programming software, it can easily check the dataset before data treatment and a great tool for data treatment and data exploration. To treat the dataset, there are 1544 data are dropped due to the 22.18% missing value in the smoking status attribute. Besides, IQR method is implemented to detect the outliers for having a better visualization of data exploration. After data treatment, it can be visualized that it is not a good dataset if it requires to train the dataset for predict the triage because the class of triage is not distributed evenly as the yellow triage is large majority (4391) among all the classes while the red triage is least minority (129). By comparing the cholesterol and age with highlighting blue as male, red as female, the data can be visualized that male have higher cholesterol level compared to female and there is no significant correlation between cholesterol and age. Hence, data treatment and data exploration for the triage dataset have been studied.

References

[1] Tam, H.L., Chung, S.F., and Lou, C.K.: 'A review of triage accuracy and future direction', BMC Emerg Med, 2018, 18, (1), pp. 58

[2] Plesner, L.L., Iversen, A.K., Langkjaer, S., Nielsen, T.L., Ostervig, R., Warming, P.E., Salam, I.A., Kristensen, M., Schou, M., Eugen-Olsen, J., Forberg, J.L., Kober, L., Rasmussen, L.S., Soletormos, G., Pedersen, B.K., and Iversen, K.: 'The formation and design of the TRIAGE study--baseline data on 6005 consecutive patients admitted to hospital from the emergency department', Scand J Trauma Resusc Emerg Med, 2015, 23, pp. 106