Part1 data exploration

In this part I’m going to explore the data set, glance through the basic information.

head()/describe()/info() are applied to view the text-information.

While a hist plot is drawn to visualize the data to get a better view.

What I learned from the exploration will later decide what pre-processing work will be done before the data is feed to the classifier

<h4>From this part we can learn that:</h4>

1. both training and testing data has no NAs, so there's no need to fill the NAs

2. some cols in training data has outliers, we need to delete them later

3. if we delete the outliers, the cols' value range will be smaller then other values' range, we need a rescaling to dealt with the problem, rescaling should be done before deleting outliers to maintain the most character of the data.

4. the label distribution of the data is imbalanced, the most label 's' has a amount of 136 while the least has a amount of 38, so oversampling should be applied to ensure we get a balanced set.