Data622_HW1_Qn2

Here is the Rpub link for the Exploratory data analysis done for both data sets
https://rpubs.com/charlsjoseph/Data622_HW1Qn2

1. Observations for dataset 1 - 'junk1.txt' file.

There are 100 observation with 3 columns(a, b , class) and dataset is balanced(means equals numbers of both classes)
There are no missing data points, no outliers and box plot shows there looks to be a relationship on feature variables v/s class variables.

With this dataset, I don't think I need more data. However I need to know more details about
1. Meta data(brief details on what does a, b and class denotes in the data set)
2. What is the end goal we are trying to achieve?

2. Observations for dataset 2 - junk2.csv

1. First of all, I see an imbalanced data set. I would go business and request for a more balanced dataset if possible.
2. Assuming that the given dataset is based on a normal distribution, the response variable is always imbalanced. So I would ask the business which performance metrics I should try to improve. Whether it is Type1 error/Precision or type2/Recall while evaluating the model performance. This is very crucial for determining the goodness of the classifier.
3. Using boxplots, we are seeing some outliers for both a, b variables. Extract those and ask the business team that if they are all genuine and determine the need of removing it from the dataset.