# DATA MINING ASSIGNMENT 1
# REPORT ON SURGICAL DATASET USING WEKA

WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

The dataset we are using to visualize in WEKA is the surgical dataset. We first convert the csv file to arff format as it is preferred by WEKA.

Loading the dataset -
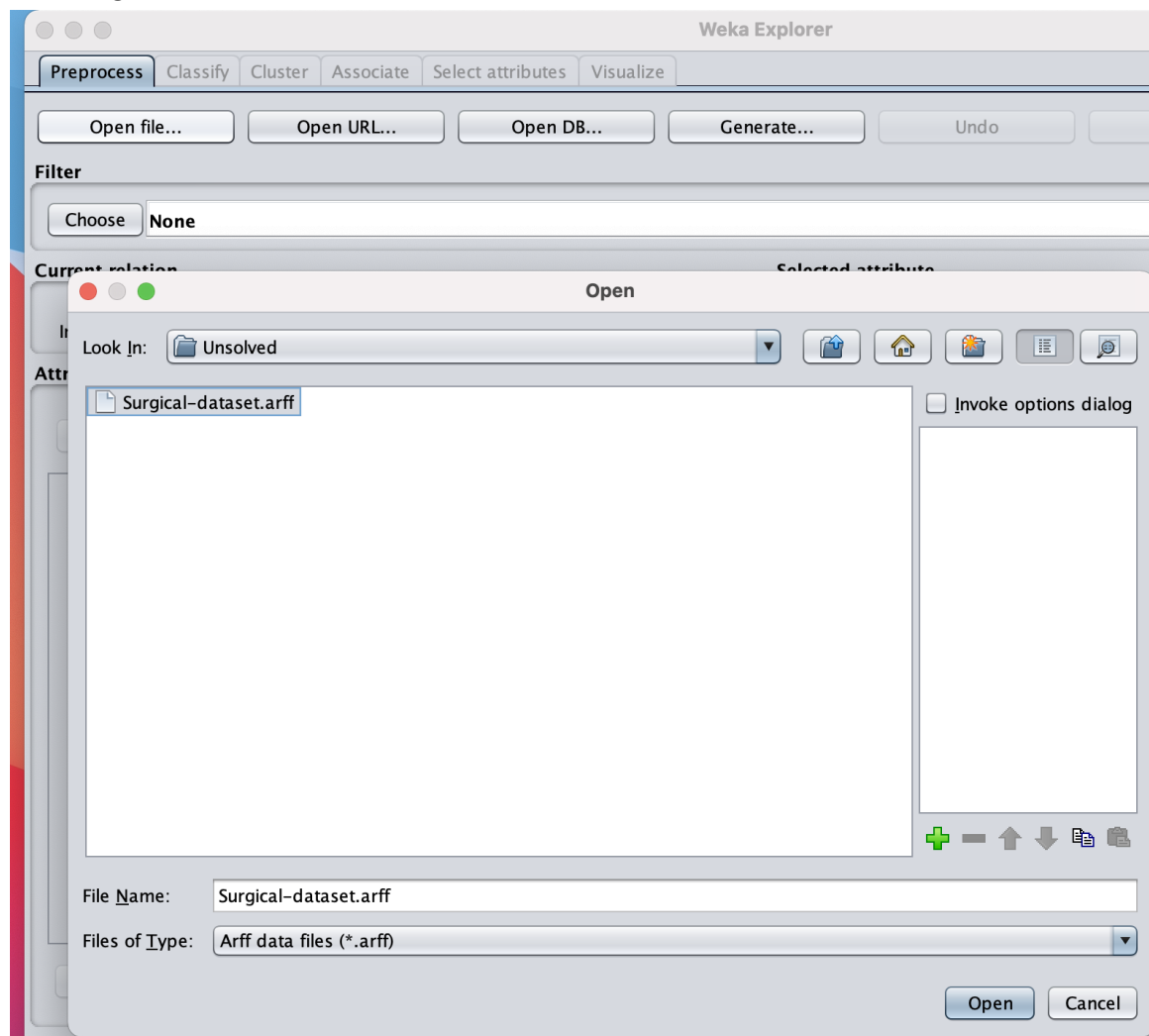
On loading the dataset into WEKA, you can see the following information on the tool.
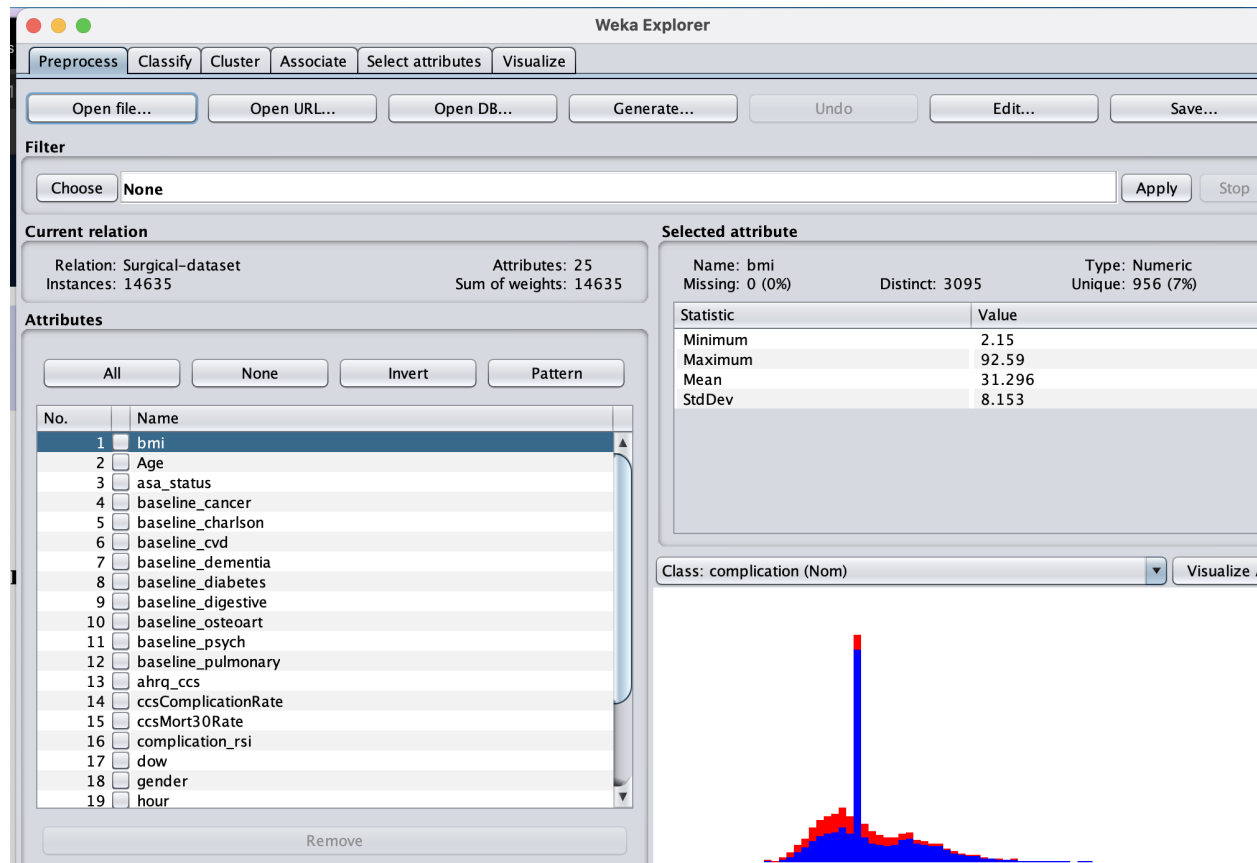

The dataset contains 25 attributes and 14635 instances. In layman terms, it consists of 25 columns and 14635 rows of data. For each selected attribute, we can see the minimum value, maximum value, mean and standard deviation of the data to the right handed side of the tool. The data is also plotted on the screen.

In the given dataset, there is no missing data for any column. If there is any missing data or null data found in any of the columns, we can actually filter out the data using WEKA's in-built filters.
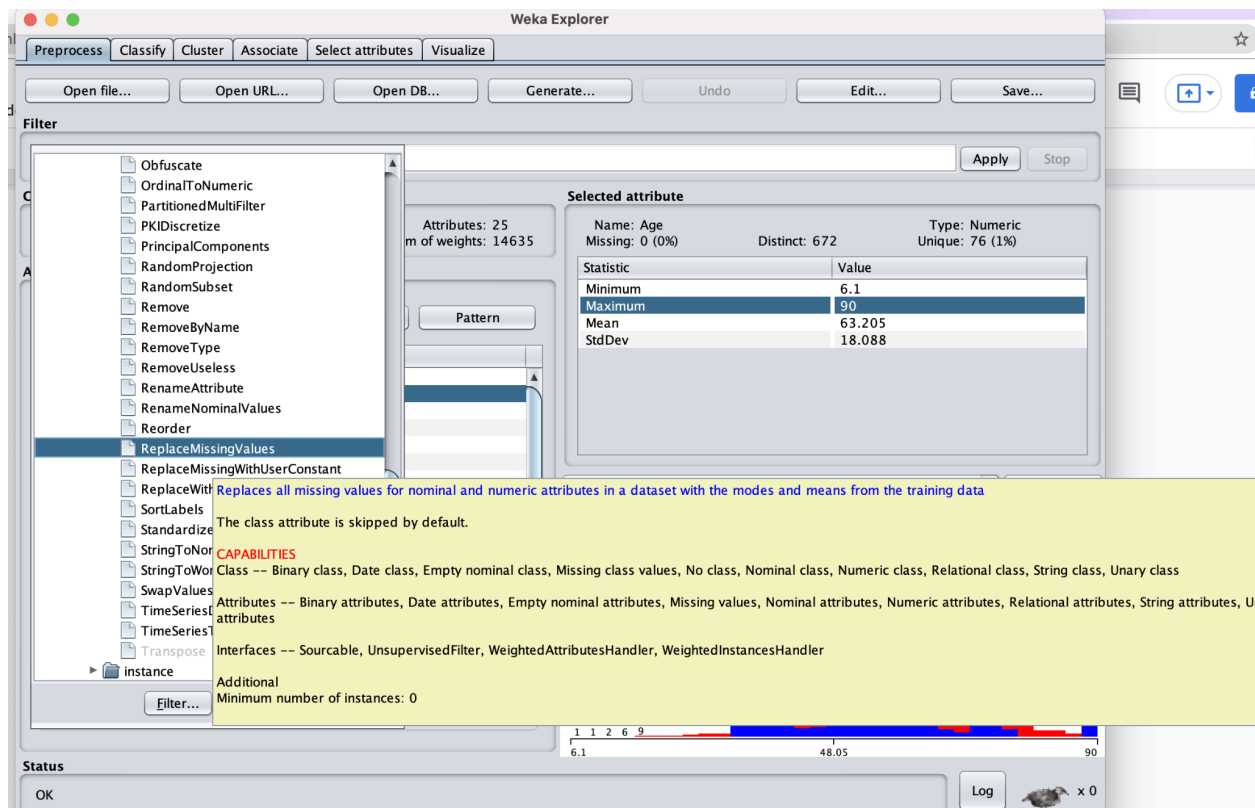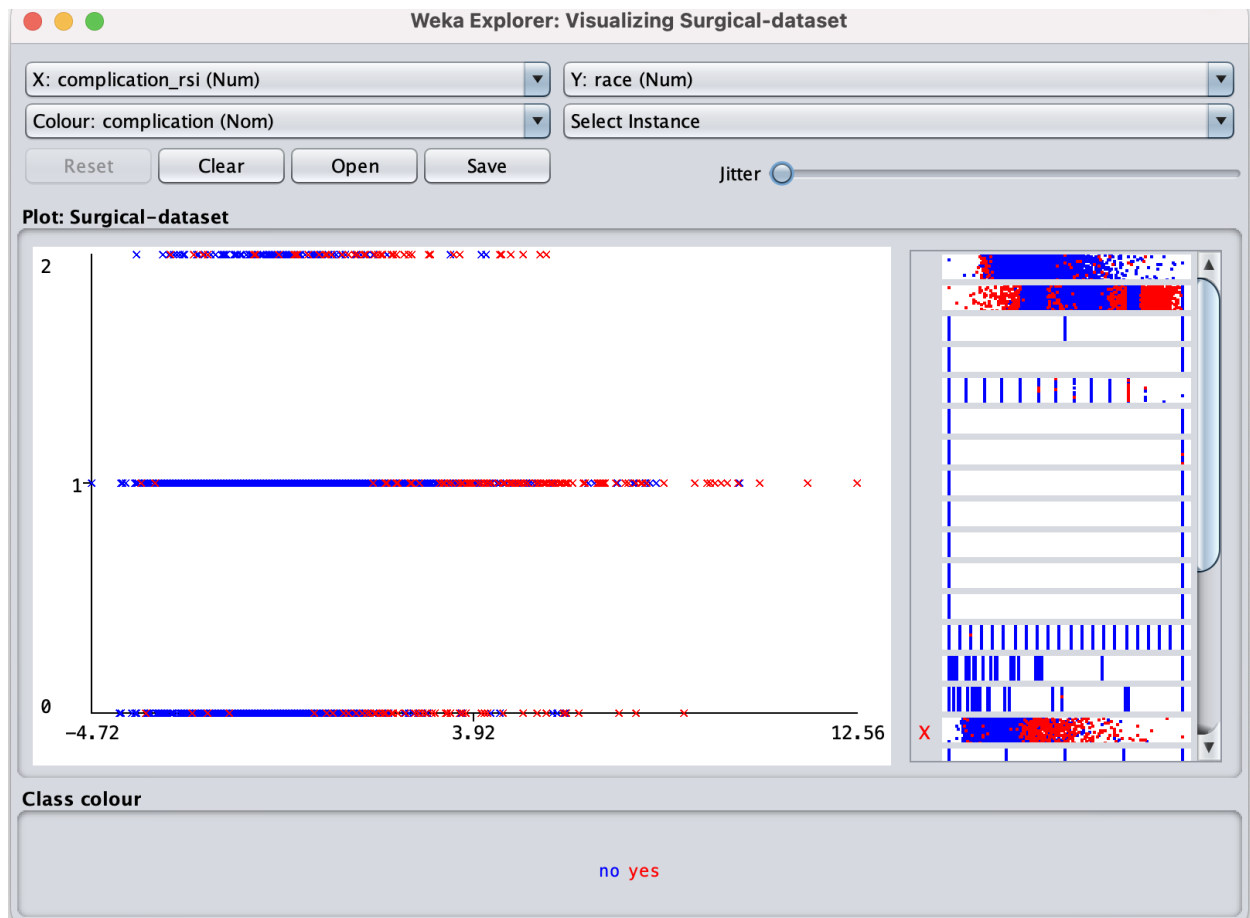
In our dataset, there is no data that is missing, hence no filtering is necessary here.

Visualization:
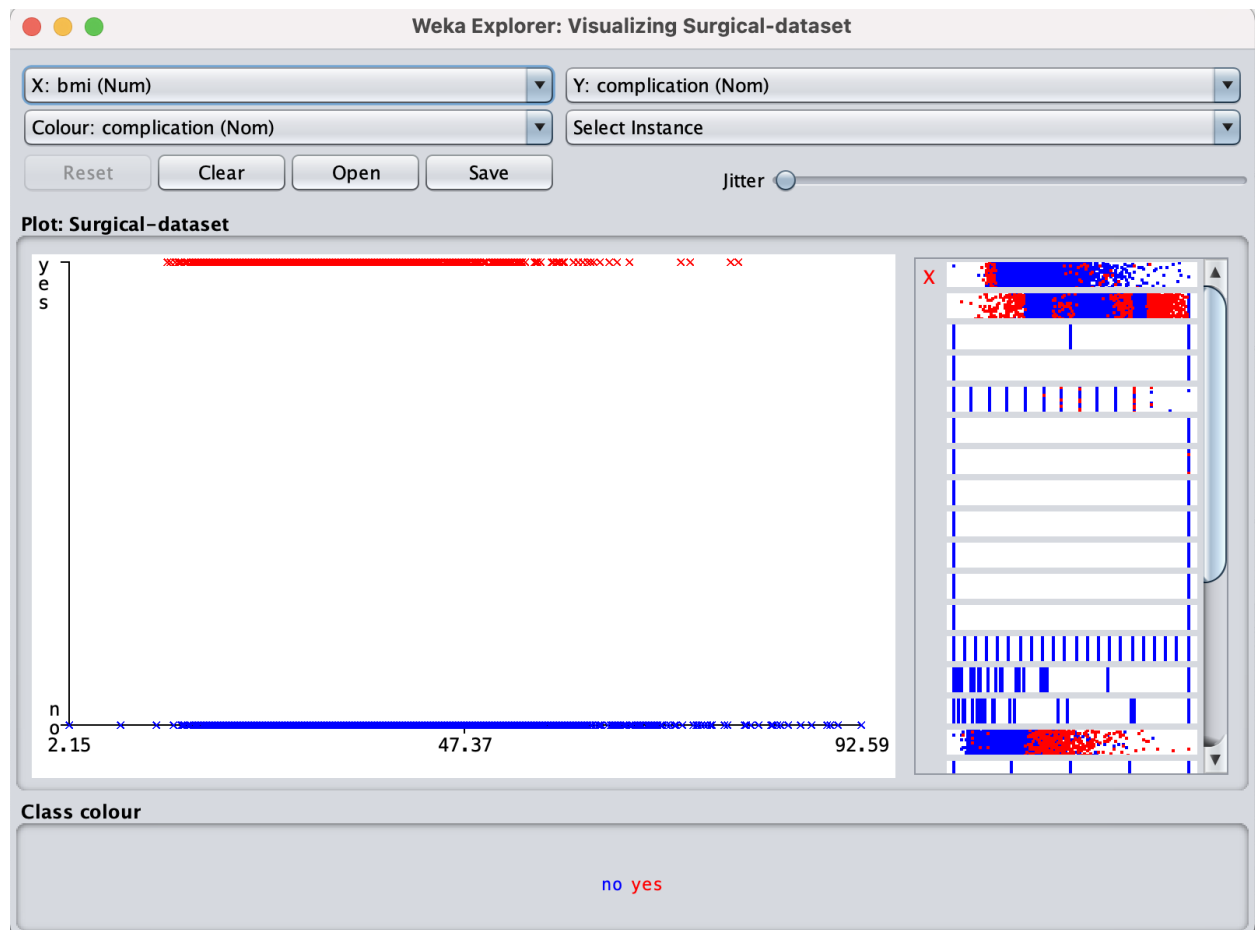
Here we have taken Age in the X-axis and Complication in the Y-axis.
The blue graph indicates that there is no complication whereas the red graph
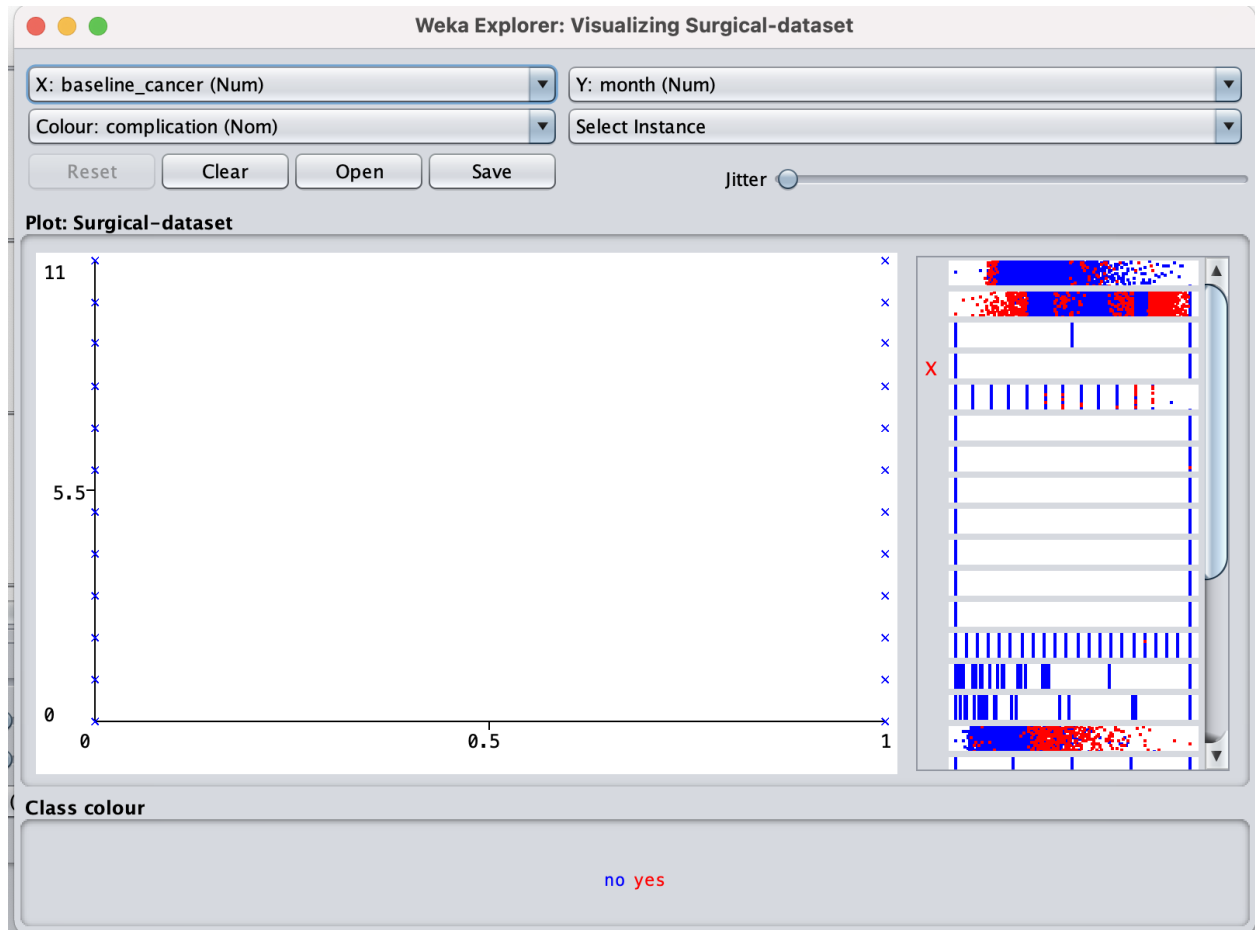indicates that is complication w.r.t age.

Here we are comparing complication_rsi with race on X-axis and Y-axis simultaneously. Here, blue points on the graph represent there is no complication rate with respect to race.

However, when we add Jitter(artificial noise) to the graph, we notice that there is complication_rsi w.r.t race for race value = 1. But for race value =0 and 2, we can see that there is not much of a change.
 Figure is as follows:
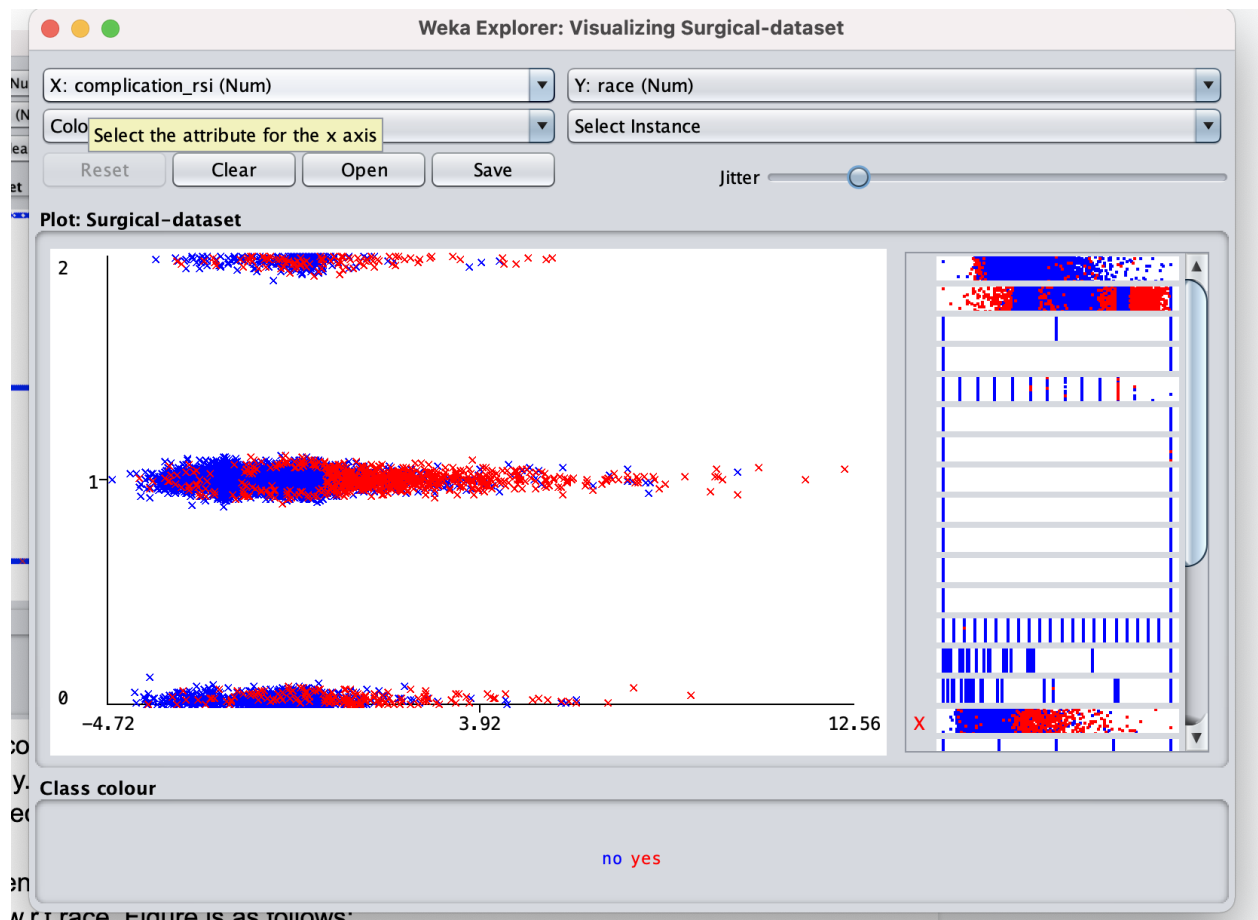
 In the next graph, we are plotting baseline_cancer on x-axis with the month on y-axis

Citations:
https://www.youtube.com/watch?v=j4tdRRT-RIg