Assignment 1

Dataset Description

The main aim is to classify breast cancer tumors as either malignant (cancerous) or benign (non-cancerous). The dataset contains features computed from digitized images of fine needle aspirate (FNA) biopsies of breast masses, and it is often used to develop and evaluate classification algorithms.
This is a binary classification problem, where the goal is to predict whether a tumor is malignant or benign based on the provided features.The two possible classes are typically labeled as malignant (M) and benign (B).

Exploartory Data Analysis

The dataset contains total of 569 rows and 32 columns. Out of which 212 data points belong to class 'M' and 357 data points belong to class 'B'. Also during exploring there were no missing data. The data is scaled using the Standard Scalar module. However there are lots of outliers present in the data which are shown below.

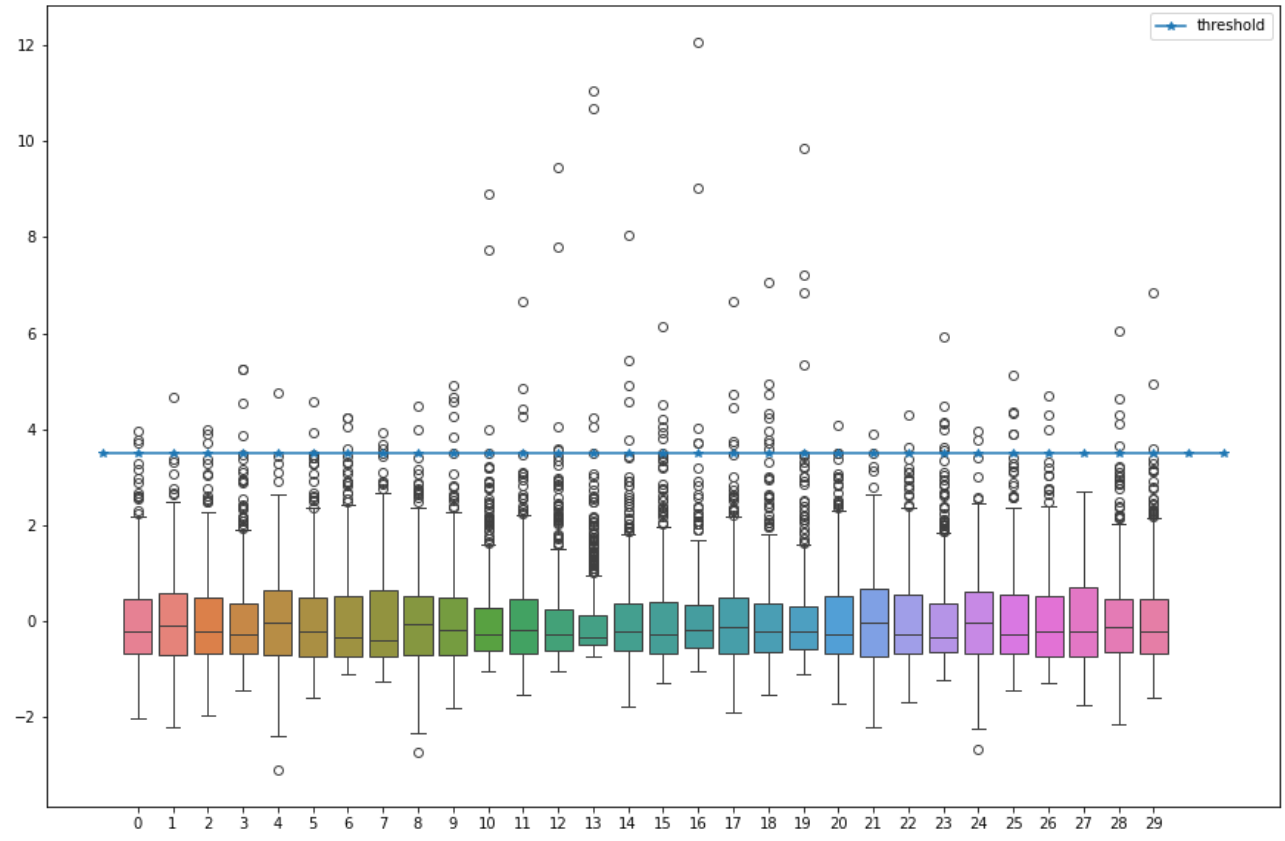

Fig. 1. Box plot for different attributes containing outliers.

In the above diagram a threshold of 3.5 is taken to separate low varianced data and high varianced data.