distribution of data after preprocesing