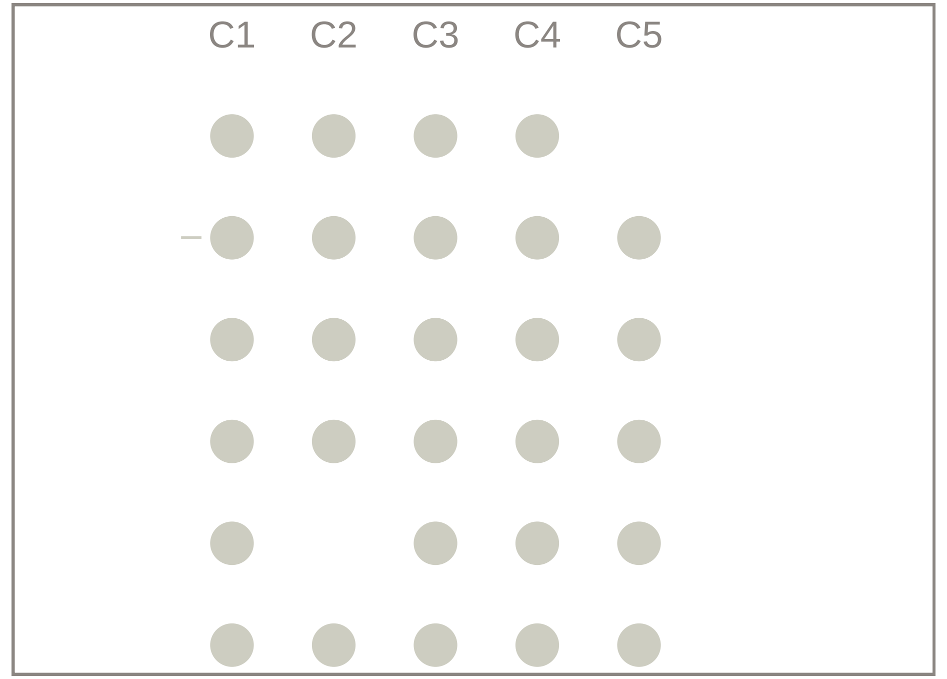
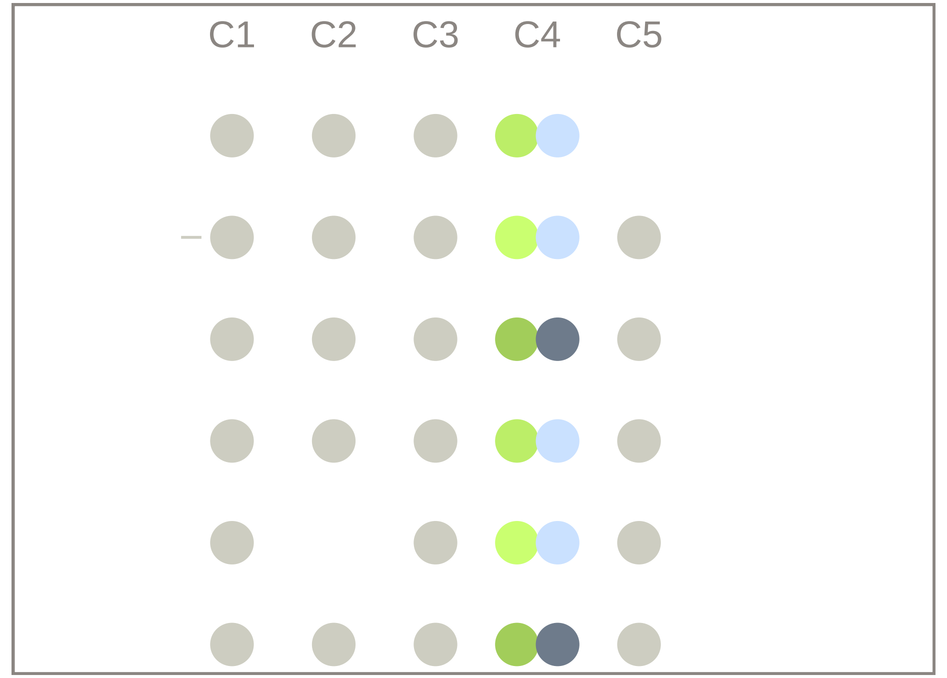
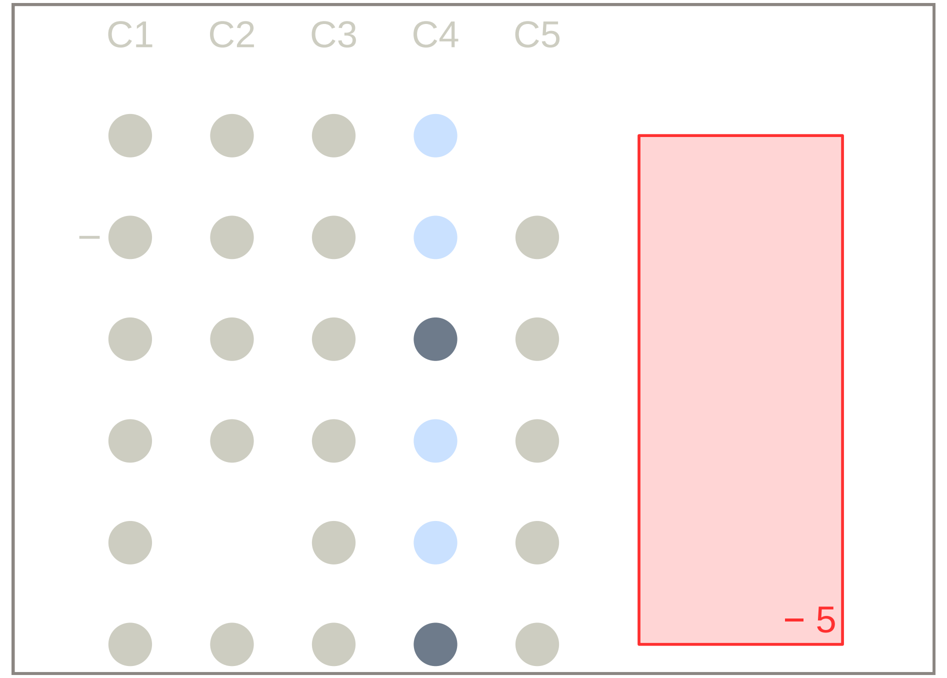


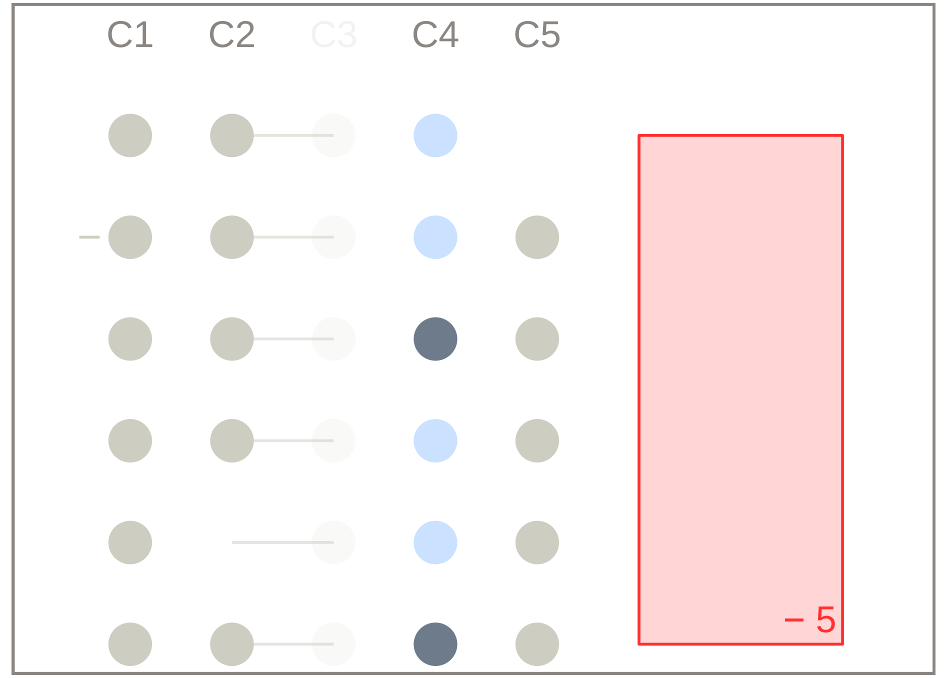
A raw data subset of the NASA metrics data program (MDP) data set.



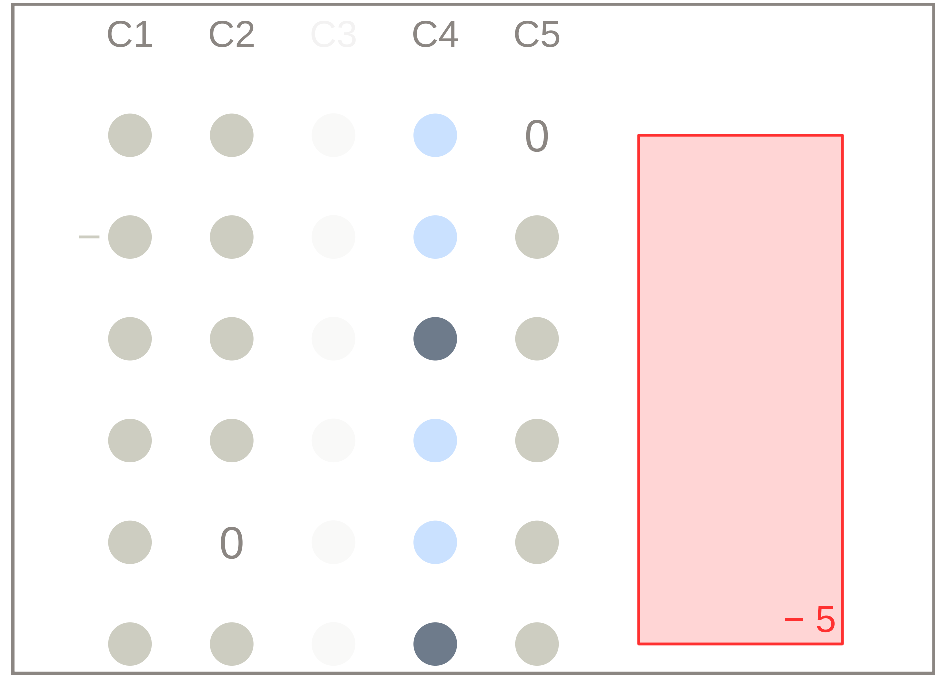
Class variable is binned, turning a continuous variable into a binary variable.



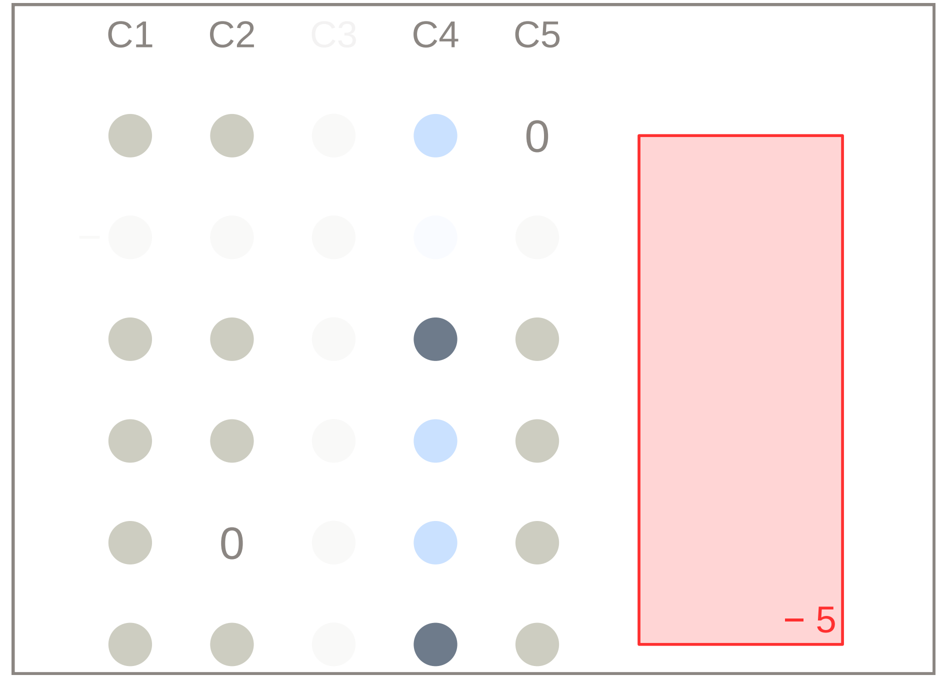
Dropped five unneeded columns. "A numeric attribute which has a constant/fixed value throughout all instances is easily identifiable as it will have a variance of zero. Such attributes contain no information with which to discern modules apart, and are at best a waste of classifier resources... This stage removes data that may be genuine, but in the context of machine learning it is of no use and is therefore discarded." (from page 551)



Remove correlated column. "In addition to constant attributes, repeated attributes occur where two or more attributes have identical values for each instance. Such attributes are therefore fully correlated, which may effectively result in a single attribute being over−represented... This stage again removes data that may be genuine, because it can be problematic when data mining." (from page 552)



"missing values have occurred because of a division by zero error... because of this we replace all missing values with zero, ensuring consistency between data sets." (from page 552)



Removal of "theoretically impossible occurrences," such as a negative year value.