Subject: Re: Sorting, counting and deleting different elements (e.g., Iodine) in a dataset
Posted by nbehrnd on Tue, 13 Sep 2022 06:55:24 GMT
View Forum Message <> Reply to Message

Dear Jon,

this is indeed an interesting question to think about further. As an early concept, equally based on the previous assignment of the Hill formula, I wrote a DW macro which subsequently uses a regular expression (regex) in a if-clause to test if the entry in question contains each of the elements (CHNO) at least once. (See the attachement below.)

Though using a macro likely eases the task (as in to offer reproducible action regardless the size of the data set, and rate of processing), there might be some obstacles ahead to extend the approach, i.e. to use multiple «filters» / «detectors» at once. To check for (CHNF), or (CHOP), or (CHOSe) as you intend is going to generate categories. This is not a problem for drawing a histogram with DW, but the syntax to probe, e.g. currently for (CHNO)

if(matchregex(MolecularFormula, "^C.*H.*N.*O.*"), "CHNO", "")

basically states

«check the regex expression on the Hill formula; if evaluated .True. return CHNO (which later may counted by DW plotting the histogram) -- else (equivalent to .False. / there is no match) return nothing».

Normally, I would try using the now empty return (above "there is no match") to nest a second test, e.g., «now test for CHOP». However, contrasting to «binning the data» as in «entries with a molecular mass, and user defined thresholds to establish categories based on this property in common»*), this approach doesn't work well enough here, because a molecule belonging to the category of (CHNO) simultaneously may belong to the category of (CHNF). So here, the discern neither is by one category in common (molecular mass), nor are the categories to probe in a relationship like (partial or complete) sub/super sets of each other.

*) DW allows to bin continuous data in preparation e.g., of a histogram; then, the bin size (e.g., interval of the molecular masses per class) applied however is uniform all across the data.

Norwid

File Attachments

1) Random_Molecules.dwar, downloaded 250 times
2) probe_CHNO.dwam, downloaded 254 times