Subject: Calculation of mean/median values in box &whisker plots.
Posted by timritchie on Wed, 24 Jul 2019 09:05:18 GMT
View Forum Message <> Reply to Message

Hello,
I have a DW file of compounds, separated into different classes, and activity data. Some compounds do not have an activity (cells are empty).
When mean values are displayed in a box or whisker plot, the values changes when the compounds with no activity are included.
I'm not sure why this should happen since they shouldn't affect the calculation of the mean.
Any thoughts on why this occurs?
Thanks,
Tim Ritchie.

Subject: Re: Calculation of mean/median values in box &whisker plots.
Posted by nbehrnd on Thu, 25 Jul 2019 08:30:43 GMT
View Forum Message <> Reply to Message

Hello Tim,

lacking a minimal working example I can only guess what you refer too. Recent work of mine with DW however equally considered data with columns lacking some entries, too. The solution working «good enough» for me, aiming for whisker plots and their statistics, however was to start with a table with each cell in the corresponding column already filled with the place holder «N/A»; to be replaced by real data only if these are at hand. (This equally may be entered prior to DW with a conditional formatting in a spred sheet, or as manual edit per cell in DW, too.) Thankfully, this kind of «other entry type» seen in other statistical programs (e.g., R) seems to be recognized by DW, too.

As an example, I populated an array about the first ten alcanes, and let DW determine their molecular weight; eventually displayed as a whisker plot (cf. alkanes_complete.dwar). In a copy of this file (alkanes_except_octaneMW.dwar) the entry about the molecular weight for octane's was substituted by the «N/A» place holder. Now, the dot about the entry missing (or, in parlance of R, about «[data] not available») is set below the others, no longer considered for further statistics -- both on screen, as well as in the plots' statistics.

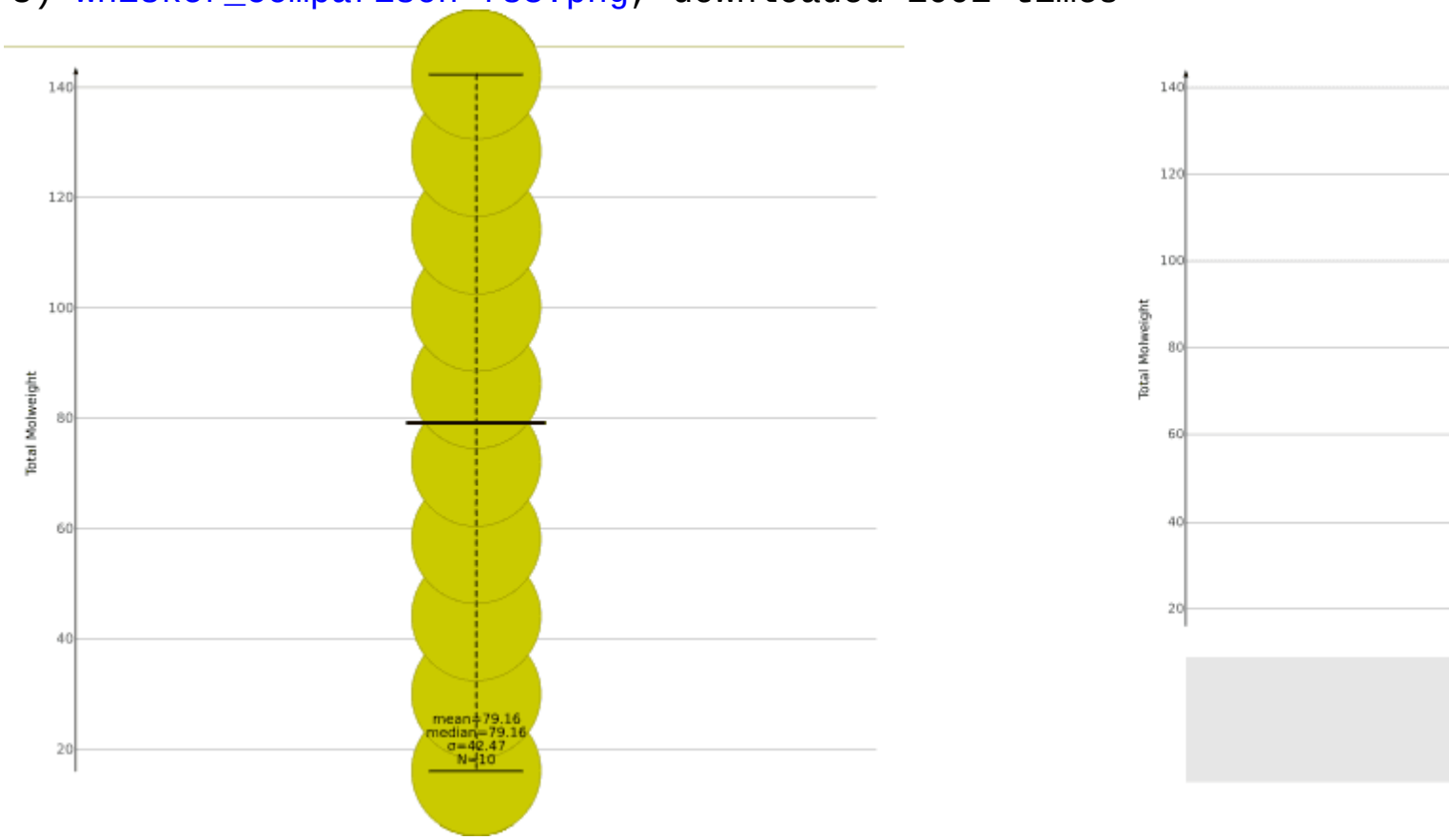
This approach was used with DataWarrior (stable release 5.0.0) running in Linux Xubuntu (18.04.2 LTS, 64 bit).

Norwid

File Attachments

1) alkanes_complete.dwar, downloaded 784 times
2) alkanes_except_octaneMW.dwar, downloaded 767 times

3) whisker_comparison-fs8.png, downloaded 1001 times



4) statistics.txt, downloaded 759 times

Subject: Re: Calculation of mean/median values in box &whisker plots.
Posted by thomas on Fri, 23 Aug 2019 22:28:12 GMT
View Forum Message <> Reply to Message

Hi Tim, I tried to reproduce, but wasn't able to do so, with both normal and logarithmic view mode on the value column. Also Norwid's alkanes_without_octaneMW.dwar does not change mean nor median, when switching off the octane. If the problem still exists, do you have a sample file? Thanks in advance and sorry for the very late reply. Thomas

---

Subject: Re: Calculation of mean/median values in box &whisker plots.
Posted by nbehrnd on Sun, 25 Aug 2019 18:13:15 GMT
View Forum Message <> Reply to Message

Hi Thomas, Hi Tim,

in my observation, all four statistical values provided in the whisker plot do change by setting manually the cell entry of molecular mass to the string of "N/A" (without the quotation marks). Processing the data

a twice allows me to retrieve the changes in the whisker plot and its statistical data, too. Here I share my approach to the task with DW (Linux version 5.0.0) with the test file alkanes_complete.dwar above:

Reading the file as-such which contains 10 complete entries:

Accessing the cell value for methane, "16.0428" is replaced by "=N/A". As expected, DW will indicate this as a non-valid entry. Of course, no whisker plot is provided now. But the dot is still present in the plot.

Next step, replacing "=N/A" by "N/A". DW accepting this now provides a Box whisker plot. The dot without associated value is sorted out, the statistical data are updated.

Two additional observations:

If "N/A" is entered for the first time, the line vanishes completely, hence shortening the list of 10 alkanes to 9 alkanes. This removal may affect other columns than the column currently worked on, too. This contradicts the aim to preserve the complete line which only has no entry for this very cell.

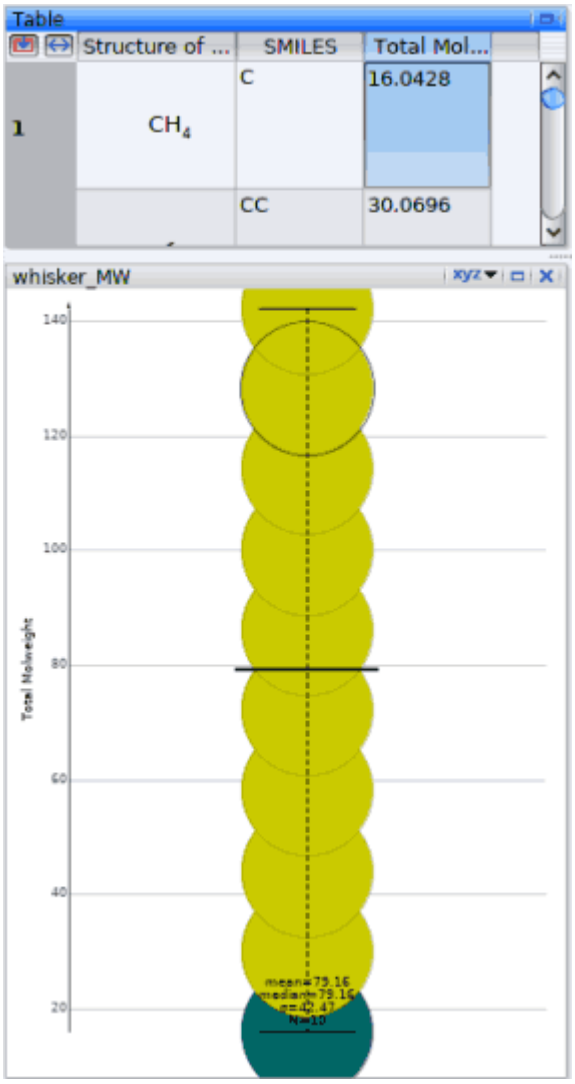
Say, there is a second entry with no value available. Then, a direct input of "N/A" via DW's edit cell function is possible without danger to loose the complete line. E.g., direct access to

Again for documentation, the file used here is attached below. Maybe there is a better way -- if so, I'm curious to learn it.

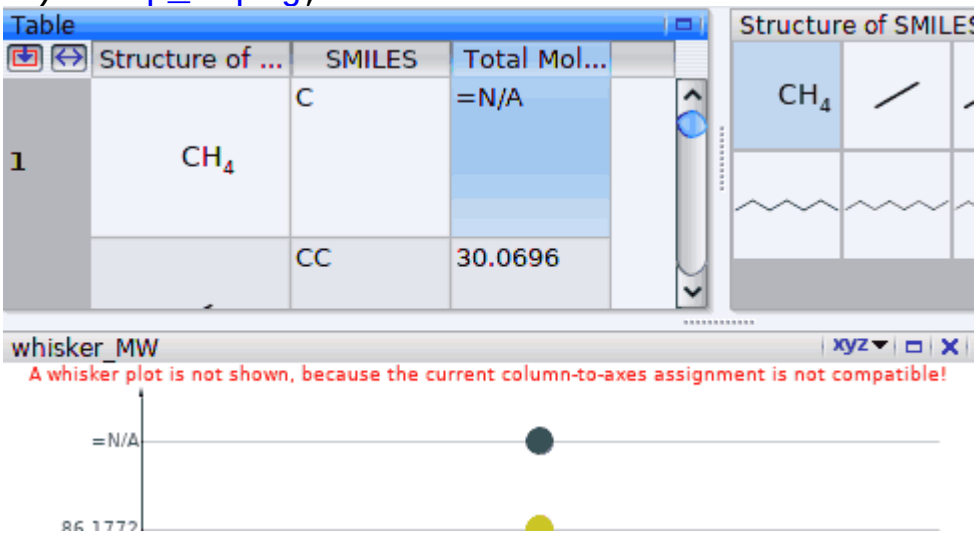
Norwid

File Attachments

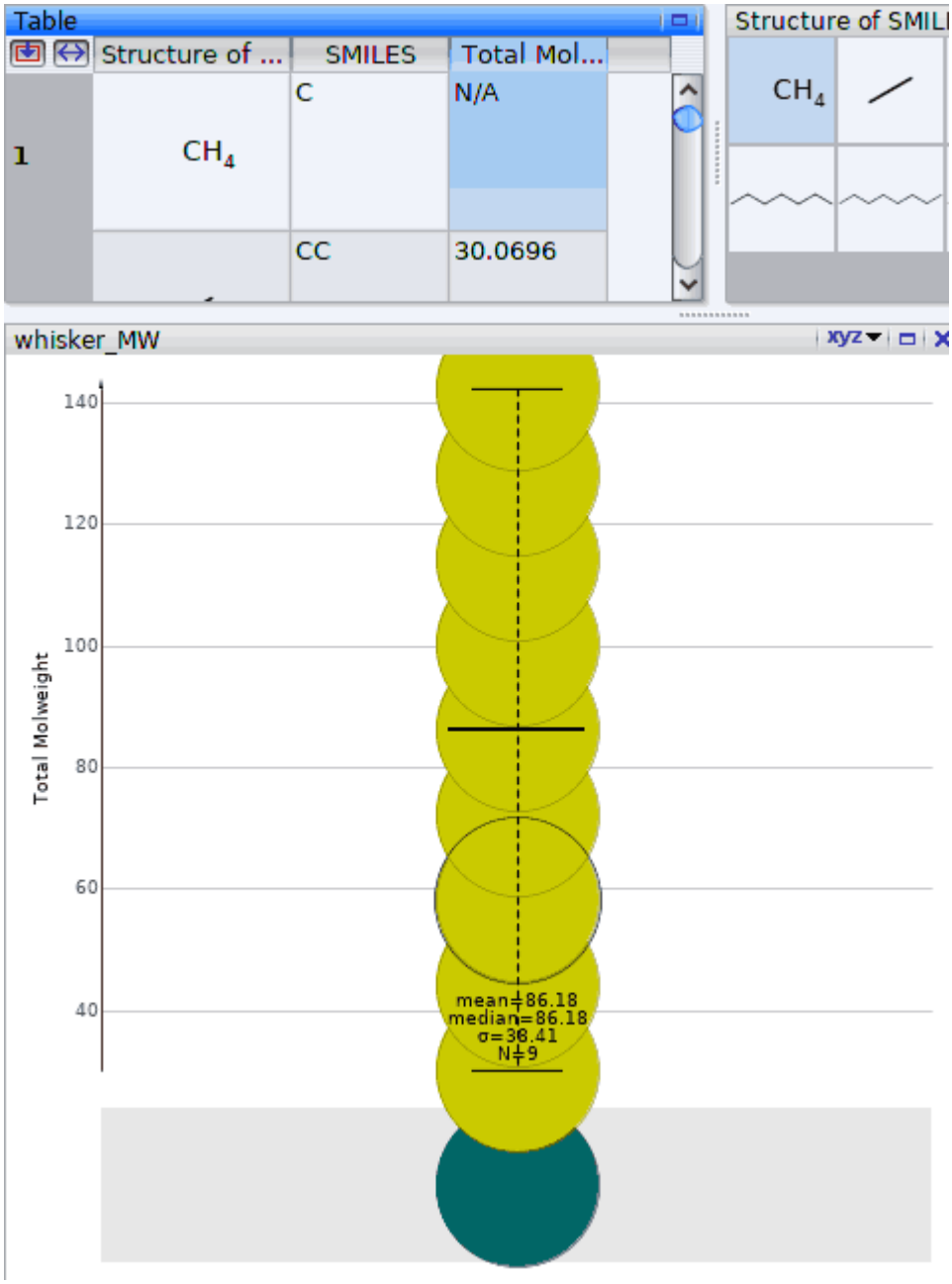
1) step_0.png, downloaded 1566 times



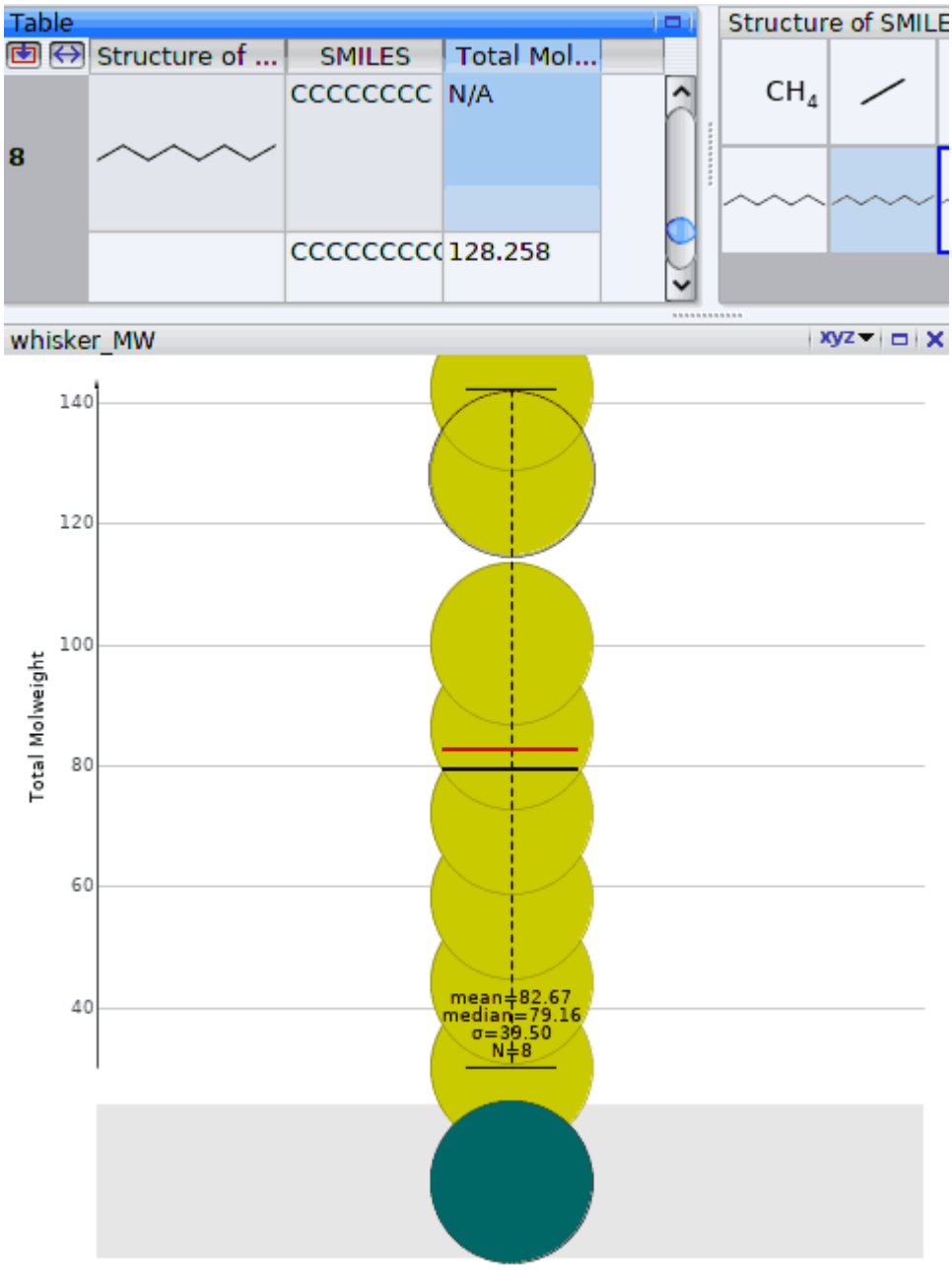
2) step_1.png, downloaded 1538 times



3) step_2.png, downloaded 1425 times



4) step_3.png, downloaded 1496 times



5) test_file.dwar, downloaded 761 times

Subject: Re: Calculation of mean/median values in box &whisker plots.
Posted by thomas on Sun, 25 Aug 2019 19:52:02 GMT
View Forum Message <> Reply to Message

Hi Norwid,

if a complete row disappears from the table or structure view after setting the numerical value in one column to 'N/A', there may be these reasons:

Potential reason 1: there is a range filter on that column. If a row is changes to not contain a numerical value anymore, then it is not in the range of the filter anymore and immediately filtered out.

Potential reason 2: the column is assigned to an axis of a graphical view and the view does not show empty values. In this case the row cannot be placed on the view anymore, because there is no value anymore. The default behavior of DataWarrior is to show in all views the same rows. Thus, if the row is removed from the graphical view, it must be removed from all other views as well, unless the graphical view is configured 'not to influence global row visibility'.

Hope this is a useful explanation,

Thomas
