Custom Statistics in dfSummary

This document shows how to customize the content of the *Stats / Values* column in data frame summaries (`summarytools::dfSummary()`) `summarytools::dfSummary()`. This feature was introduced in version 1.0.0, July 2021.

This feature request came up several times in a form or another, mostly on GitHub.

How it works

Two new options were created: `dfSummary.custom.1` and `dfSummary.custom.2`. The first one has a predefined value – it is the one that makes up the fourth row of the cell (showing IQR and CV). The second one is set to `NA` by default. If both options are defined (non-`NA`), the cell will now span on 5 lines rather than 4, provided there are no additional line breaks occurring.

Baseline

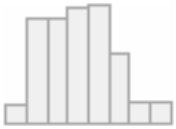
We'll use the first column of *iris* for this demo. So let's see the results as they are before making any changes.

First let's set things up:

```
library(knitr)
opts_chunk$set(comment = NA,
               prompt  = FALSE,
               cache   = FALSE,
               echo    = TRUE,
               results = 'asis')
#knit_theme$set("edit-kwrite")
# https://rclickhandbuch.files.wordpress.com/2014/09/knitrthemesoverview.pdf
library(summarytools)
st_options(plain.ascii     = FALSE,
           headings        = FALSE,
           footnote        = NA,
           round.digits    = 1,
           dfSummary.varnumbers   = FALSE,
           dfSummary.valid.col    = FALSE,
           dfSummary.silent       = TRUE,
           dfSummary.style        = "grid",
           tmp.img.dir            = "img")
```

And then show the baseline:

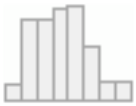
```
iris_subset <- iris[1]
dfSummary(iris_subset, graph.magnif = .45)
```

Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
Sepal.Length [numeric]	Mean (sd) : 5.8 (0.8) min < med < max: 4.3 < 5.8 < 7.9 IQR (CV) : 1.3 (0.1)	35 distinct values		0 (0.0%)

Example 1 - Removing the IQR (CV) line

Setting the first option to `NA` will do just that:

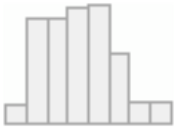
```
st_options(dfSummary.custom.1 = NA)
dfSummary(iris_subset, graph.magnif = .35)
```

Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
Sepal.Length [numeric]	Mean (sd) : 5.8 (0.8) min < med < max: 4.3 < 5.8 < 7.9	35 distinct values		0 (0.0%)

Example 2 : Adding Q1 & Q3

Here we're going to create the expression that is needed to generate the statistics we want; since this bit of is going to be interpreted while looping on column data, there are some variables that are available to us. The most important is, well, `column_data`. Another one that you might want to use is `round.digits`; we have set it to `1` in the initial chunk.

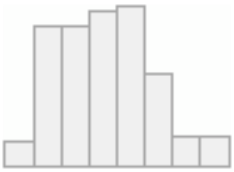
```
st_options(
  dfSummary.custom.1 =
    expression(
      paste(
        "Q1 - Q3 :",
        round(
          quantile(column_data,
                   probs = .25,
                   type = 2,
                   names = FALSE,
                   na.rm = TRUE),
          digits = 1
        ), " - ",
        round(
          quantile(column_data,
                   probs = .75,
                   type = 2,
                   names = FALSE,
                   na.rm = TRUE),
          digits = 1
        )
      )
    )
)
dfSummary(iris_subset, graph.magnif = .45)
```

Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
Sepal.Length [numeric]	Mean (sd) : 5.8 (0.8) min < med < max: 4.3 < 5.8 < 7.9 Q1 - Q3 : 5.1 - 6.4	35 distinct values		0 (0.0%)

Example 3: Inserting back the IQR & CV

It is always possible to revert the first custom stat to its initial value by using `st_options(dfSummary.custom.1 = "default")`. But let's make things a bit more interesting by actually showing these under the Q1 & Q3 line that we have just defined.

```
st_options(
  dfSummary.custom.2 =
    expression(
      paste(
        paste0(
          trs("iqr"), " (", trs("cv"), ") : "
        ),
        format_number(
          IQR(column_data, na.rm = TRUE),
          round.digits
        ),
        " (",
        format_number(
          sd(column_data, na.rm = TRUE) /
              mean(column_data, na.rm = TRUE),
          round.digits
        ),
        ")",
        collapse = "",
        sep = ""
    )
  )
)
dfSummary(iris_subset, graph.magnif = .65)
```

Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
Sepal.Length [numeric]	Mean (sd) : 5.8 (0.8) min < med < max: 4.3 < 5.8 < 7.9 Q1 - Q3 : 5.1 - 6.4 IQR (CV) : 1.3 (0.1)	35 distinct values		0 (0.0%)

Don't forget to set `na.rm = TRUE` whenever necessary. Otherwise, just use your imagination!

Useful links:

1. Introduction to summarytools (package vignette)
2. Summarytools in R Markdown Documents (package vignette)
3. Data Frame Summaries in PDF's (supplemental documentation)