Data Frame Summaries in PDF's

Dominic Comtois

2020-12-29

Yes, at last. It's not perfect but it's workable. I put this off for a long time, as I thought it would absolutely require a *Pandoc Lua filter*, and I was just too busy with other things. As I learn a bit more about LaTeX, I now realize that a "simple" `\renewcommand` does the trick.

So here it is, starting with the *YAML* section.

I. YAML Header

There is a *tex* file to include. For the *xelatex* engine, it's not mandatory, but there are several advantages to it, and I now use it systematically.

```
---
title: "Data Frame Summaries in PDF's"
output:
  pdf_document:
    latex_engine: xelatex
    includes:
      in_header: ./fig-valign.tex
---
```

II. Included Preamble *Tex* File

This is the LaTeXcontent that you'll need to copy in your own *fig-align.tex*:

```
\usepackage{graphicx}
\usepackage[export]{adjustbox}
\usepackage{letltxmacro}
\LetLtxMacro{\OldIncludegraphics}{\includegraphics}
\renewcommand{\includegraphics}[2][]{\raisebox{0.5\height}%
  {\OldIncludegraphics[valign=t,#1]{#2}}}
```

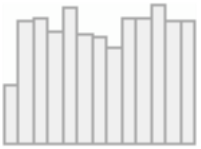
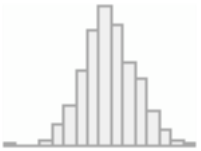
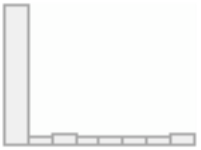
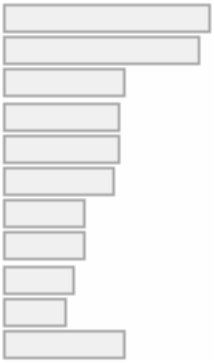
III. R Code

```
library(summarytools)
st_options(
  plain.ascii = FALSE,
  style = "rmarkdown",
  dfSummary.style = "grid",
  dfSummary.valid.col = FALSE,
  dfSummary.graph.magnif = .52,
  tmp.img.dir = "/tmp"
)
```

```
define_keywords(title.dfSummary = "Data Frame Summary in PDF Format")
dfSummary(tobacco)
```

Data Frame Summary in PDF Format

tobacco Dimensions: 1000 x 9
Duplicates: 2

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	gender [factor]	1. F 2. M	489 (50.0%) 489 (50.0%)		22 (2.2%)
2	age [numeric]	Mean (sd) : 49.6 (18.3) min < med < max: 18 < 50 < 80 IQR (CV) : 32 (0.4)	63 distinct values		25 (2.5%)
3	age.gr [factor]	1. 18-34 2. 35-50 3. 51-70 4. 71 +	258 (26.5%) 241 (24.7%) 317 (32.5%) 159 (16.3%)		25 (2.5%)
4	BMI [numeric]	Mean (sd) : 25.7 (4.5) min < med < max: 8.8 < 25.6 < 39.4 IQR (CV) : 5.7 (0.2)	974 distinct values		26 (2.6%)
5	smoker [factor]	1. Yes 2. No	298 (29.8%) 702 (70.2%)		0 (0.0%)
6	cigs.per.day [numeric]	Mean (sd) : 6.8 (11.9) min < med < max: 0 < 0 < 40 IQR (CV) : 11 (1.8)	37 distinct values		35 (3.5%)
7	diseased [factor]	1. Yes 2. No	224 (22.4%) 776 (77.6%)		0 (0.0%)
8	disease [character]	1. Hypertension 2. Cancer 3. Cholesterol 4. Heart 5. Pulmonary 6. Musculoskeletal 7. Diabetes 8. Hearing 9. Digestive 10. Hypotension [3 others]	36 (16.2%) 34 (15.3%) 21 (9.5%) 20 (9.0%) 20 (9.0%) 19 (8.6%) 14 (6.3%) 14 (6.3%) 12 (5.4%) 11 (5.0%) 21 (9.5%)		778 (77.8%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
9	samp.wgts [numeric]	Mean (sd) : 1 (0.1) min < med < max: 0.9 < 1 < 1.1 IQR (CV) : 0.2 (0.1)	0.86!: 267 (26.7%) 1.04!: 249 (24.9%) 1.05!: 324 (32.4%) 1.06!: 160 (16.0%) ! rounded		0 (0.0%)

Closing Remarks

Since we redefined the command `includegraphics`, all images included using `[](some-image.png)` will be impacted. In some cases this will likely be problematic. Eventually we will find a more robust solution without such undesired side-effects. If you are well versed in LaTeXand think you can solve this, by all means get in touch with me.