visdat: Visualising Whole Data Frames

Nicholas Tierney[1]

[1]Monash University

01 August 2017

Paper DOI: http://dx.doi.org/10.21105/joss.00355
Software Repository: https://github.com/ropensci/visdat
Software Archive: http://dx.doi.org/10.5281/zenodo.845960

Summary

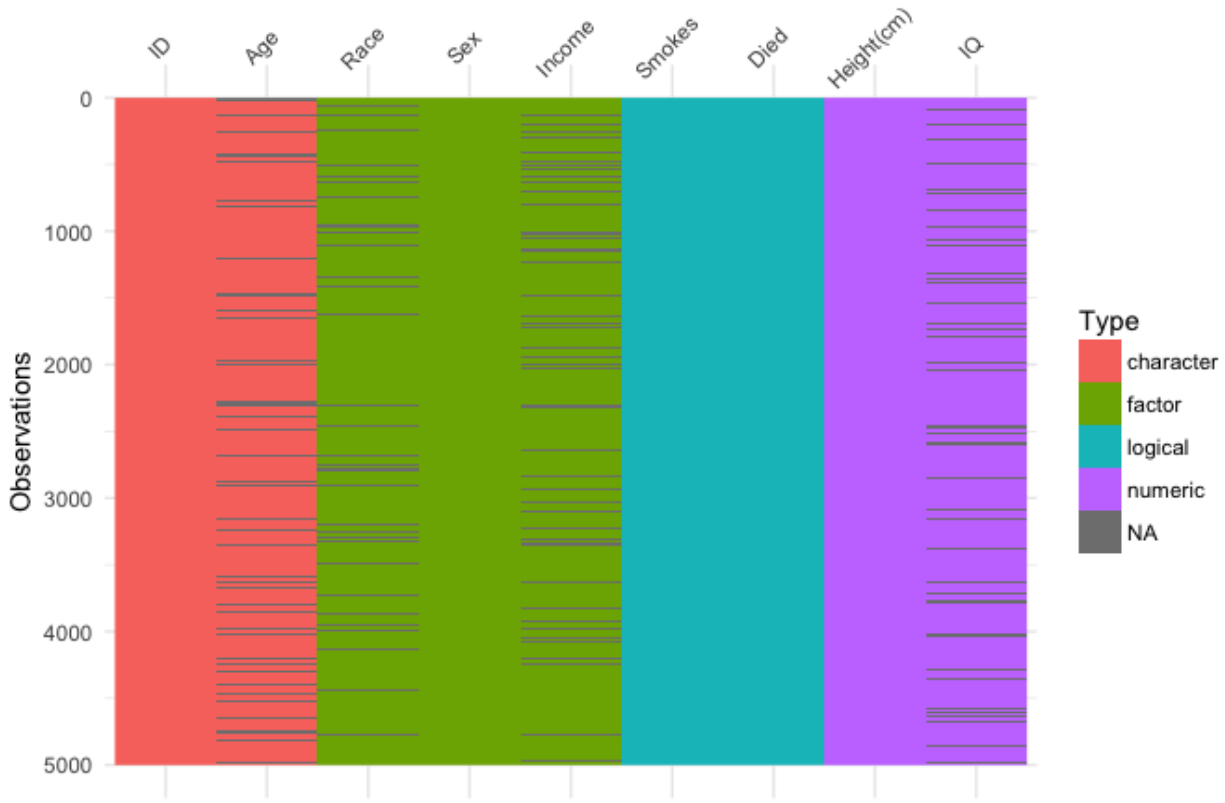
When you receive a new dataset you need to look at the data to get a sense of what is in it, and understand potential problems and challenges to get it analysis-ready. "Taking a look at the data" can mean different things. For example: examining statistical summaries (minimum, maximum, mean, inter-quartile range), finding missing values, checking data formatting, creating graphical summaries such as histograms, scatter plots, box plots, and more.

When handling typical real-world data, these preliminary exploratory steps can become difficult to perform when values are not what you expect. For example, income might be a factor instead of numeric, date could be a number not a character string or a date class, or values could be missing when they shouldn't be. Often times, you discover that you had expectations of the data, which are hard to realise until they are a problem. This is similar to how one might not think to buy more light bulbs until one goes out: when you use data in an exploratory scatter plot, or a preliminary model, you often don't realise your data is in the wrong format until that moment. What is needed is a birds eye view of the data, which tells you what classes are in the dataframe, and where the missing data are.

`visdat` is an R (R Core Team 2016) package that provides a tool to "get a look at the data" by creating heatmap-like visualisations of an entire dataframe, which provides information on: classes in the data, missing values, and also comparisons between two datasets. `visdat` takes inspiration from `csv-fingerprint`, and is powered by ggplot2 (Wickham, Chang, and RStudio 2016), which provides a consistent, powerful framework for visualisations that can be extended if needed.

Plots are presented in an intuitive way, reading top down, just like your data. Below is a plot using `vis_dat()` of some typical data containing missing values and data of a variety of classes.

```
library(visdat)
vis_dat(typical_data)
```



`visdat` will continue to be improved over time, to improve speed in computation and improve interactive plotting.

Acknowlegements

I would like to thank the two reviewers, Mara Averick and Sean Hughes, for their helpful suggestions that resulted in a much better package, and rOpenSci for providing the support of the onboarding package review that facilitated these improvements.

References

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Winston Chang, and RStudio. 2016. *Ggplot2: An Implementation of the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.