The Journal of Open Source Software

ggmatplot: An R package for data visualization on wide-format data

Xuan Liang[1], Francis K. C. Hui[1], Dilinie Seimon[2], and Emi Tanaka[2]

1 Research School of Finance, Actuarial Studies and Statistics, The Australian National University **2** Department of Econometrics and Business Statistics, Monash University

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

The layered grammar of graphics (H. Wickham, 2010), implemented as the ggplot2 package (Hadley Wickham, 2016) in the statistical language R (R Core Team, 2021), is a powerful and popular tool to create versatile statistical graphics. This graphical system, however, requires input data to be organised in a manner that a data column is mapped to an aesthetic element (e.g. x-coordinate, y-coordinate, color, size), which create friction in constructing plots with an aesthetic element that span multiple columns in the original data by requiring users to re-organise the data.

The ggmatplot, built upon ggplot2, is an R-package that allows quick plotting across the columns of matrices or data with the result returned as a ggplot object. The package is inspired by the function matplot() in the core R graphics system, thus ggmatplot can be considered as a ggplot version of matplot with the benefits of customising the plots as any other ggplot objects via ggplot2 functions.

Statement of need

Input data to construct plots with ggplot2 require data to be organised in a manner that maps data columns to aesthetic elements. This generally works well where data is tidied in a rectangular form, referred to as “tidy data” (Hadley Wickham, 2014), where each row represents an observational unit, each column represents a variable, and each cell represents a value. In some cases, what constitutes a variable (or observational unit), hence a column (or row), in a tidy data can be dependent upon interpretation or downstream interest (e.g. Tables 1 and 2 can be both considered as tidy data), but a clear violation of tidy data principles is when the column names contain data values, e.g. Table 3 contain the name of the species across a number of column names.

Table 1: Restaurant rating data in "tidy" form. The first column shows the restaurant ID, and the next four columns show the average ratings (out of 5) for food, service, ambience and overall, respectively.

	Rating			
Restaurant	Food	Service	Ambience	Overall
R1	4	3	4	4
R2	4	5	4	4
R3	3	4	5	3
R4	2	4	4	3
R5	3	4	4	3

Table 2: Another form for the restaurant rating data in Table 1. In @Wickham2014-gy, this format is called the ”molten” data.

Restauant	Rating type	Rating
R1	food	4
R1	service	3
R1	ambience	4
R1	overall	4
R2	food	4
R2	service	5
R2	ambience	4
R2	overall	4
R3	food	3
R3	service	4
R3	ambience	5
R3	overall	3
R4	food	2
R4	service	4
R4	ambience	4
R4	overall	3
R5	food	3
R5	service	4
R5	ambience	4
R5	overall	3

Table 3: Spider abundance data with environmental covariates. The rows correspond to the site, the first two columns are environmental covariates that measure the soil dry mass and cover moss, and the following five columns shows the abundance of the species.

	Environment covariates		Species abundance				
Site	Soil dry mass	Moss	Alopcune	Arctlute	Pardpull	Trocterr	Zoraspin
1	2.3321	3.0445	10	0	45	57	4
2	3.0493	1.0986	2	0	37	65	9
3	2.5572	2.3979	20	0	45	66	1

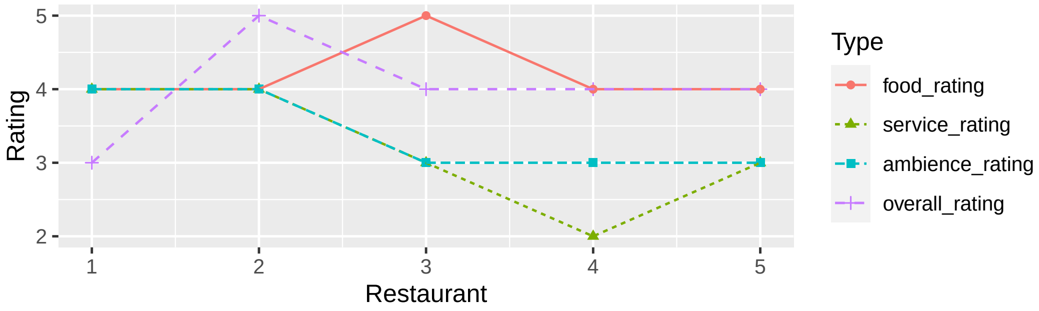
The organisation of the data is largely dependent on the downstream analysis and there is no one correct way to do this. Some forms of multivariate data, e.g. Table 3, are prevalent in the field because it aligns as an input data for a modelling software and/or the format is more convenient for input or view of the data in spreadsheet format. However, this format is not consistent with the required format for ggplot2, and consequently, plotting with ggplot2 interrupts the workflow of a user that is trying to quickly visualise these types of data. The ggmatplot R-package provides a solution to this common friction in producing plots with ggplot2.

Examples

In this section we demonstrate the use of the ggmatplot package and contrast the specification with ggplot2 with data wrangling using dplyr and tidyr (Hadley Wickham et al., 2019) using the example data in Tables 1 and 3, which are stored in the objects wide_df and abun_df, respectively.

Example 1

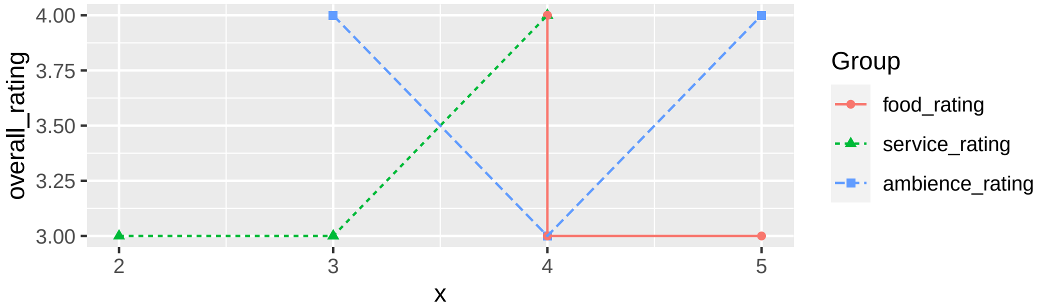
```r
library(ggmatplot)
ggmatplot(x = wide_df[, -1], plot_type = "both",
          xlab = "Restaurant",  ylab = "Rating", legend_title = "Type")
```



```r
library(ggplot2)
library(tidyr) # or library(tidyverse)
wide_df %>%
  pivot_longer(contains("rating"),
               names_to = "rating_type",
               values_to = "rating") %>%
  ggplot(aes(restaurant, rating, color = rating_type)) +
  geom_point() +
  geom_line(aes(group = rating_type,
                linetype = rating_type))
```

Example 2

```r
ggmatplot(x = wide_df[, 2:4], y = wide_df[, 5], plot_type = "both")
```



Discussion

The `ggmatplot` R-package provides a solution to a common friction to quickly plotting multivariate data where the primary interest is mapping the column names as an aesthetic element. The solution provided however is a recipe-driven approach where the user can only produce plot types as many there are included in the `plot_type` option. Future development of the package could benefit from using a grammar approach, like in Wilkinson (2005) and H. Wickham (2010), where plot types can be extensible.

Acknowledgements

FKCH was supported by ARC DECRA XXX.

References

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Wickham, H. (2010). A layered grammar of graphics. *Journal of computational and graphical statistics: a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*.

Wickham, Hadley. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23. doi:10.18637/jss.v059.i10

Wickham, Hadley. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, Hadley, Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686

Wilkinson, L. (2005). *The grammar of graphics*. Springer.