




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

Xuan Liang¹, Francis K. C. Hui¹, Dilinie Seimon², and Emi Tanaka²

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:

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).

Summary

The layered grammar of graphics (Wickham, 2010), implemented as the `ggplot2` package (Wickham, 2016) in the statistical language R (R Core Team, 2021), is a powerful and popular tool to create versatile statistical graphics. However, this graphical system 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 creates friction in constructing plots with an aesthetic element that span multiple columns in the original data by requiring users to re-organise the data.

Regarding Francis' question, I don't know if the long format or wide format data are well defined in the literature. If it is, we definitely can mention it. If not, do we need to define it? –commented by XL

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 – as such, `ggmatplot` may be considered as a `ggplot` version of `matplot` with the benefits of customising the plots as any other `ggplot` objects via `ggplot2` functions, as well as offering several other plotting types that are not immediately available from `matplot` directly, such as comparative violin plots.

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 long rectangular form, often referred to as “tidy data” (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), and hence a column (or row), in 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*. THIS IS WRONG? THERE ARE NO SPECIES HERE?

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 many scientific fields because it aligns with the input data for a particular modelling software, and/or the format is more convenient for input or view of the data in spreadsheet format (say). Unfortunately, this format is not consistent with the required format for `ggplot2`, and consequently plotting with `ggplot2` interrupts the workflow of

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.

Restaurant	Average rating			
	Food	Service	Ambience	Overall
R1	4.3	3.4	4.3	4.9
R2	4.3	5.0	4.5	4.4
R3	3.2	4.4	5.0	3.0
R4	2.3	4.6	4.4	3.8
R5	3.9	4.8	4.2	3.3

Table 2: Another form for the restaurant rating data in Table 1. In Wickham (2014), this format is called the "molten" data.

Restauant	Rating type	Average rating
R1	food	4.3
R1	service	3.4
R1	ambience	4.3
R1	overall	4.9
R2	food	4.3
R2	service	5.0
R2	ambience	4.5
R2	overall	4.4
R3	food	3.2
R3	service	4.4
R3	ambience	5.0
R3	overall	3.0
R4	food	2.3
R4	service	4.6
R4	ambience	4.4
R4	overall	3.8
R5	food	3.9
R5	service	4.8
R5	ambience	4.2
R5	overall	3.3

Table 3: The first 6 rows and 11 columns of the snowfall data for Grand Rapids, Michigan in the R pacakge extttmosaicData (Prium, Kaplan & Horton, 2021).

SeasonStart	SeasonEnd	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar
1893	1894	0	0	0	0.0	8.0	24.9	12.5	6.8	4.8
1894	1895	0	0	0	0.0	7.5	5.3	21.5	8.0	22.5
1895	1896	0	0	0	0.4	23.2	15.0	.	8.5	2.0
1896	1897	0	0	0	0.2	8.0	8.0	4.9	11.2	12.0
1897	1898	0	0	0	0.0	1.4	8.0	15.5	29.5	0.0
1898	1899	0	0	0	0.0	18.5	18.0	20	3.4	16.0

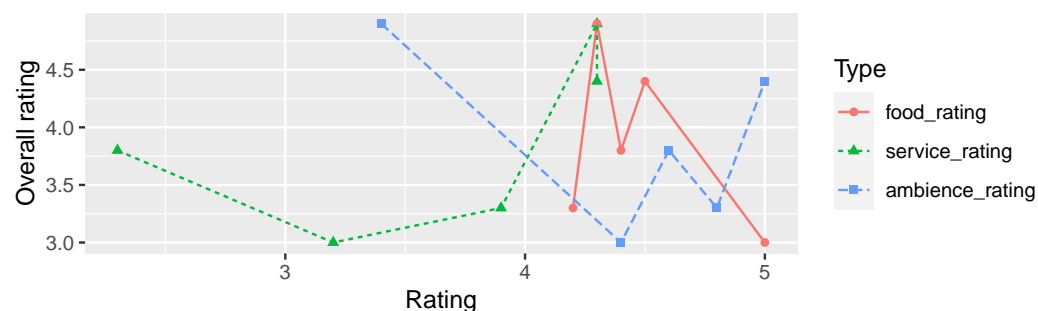


Figure 1: Line plot of the food, service and ambience rating against the overall restaurant rating.

a user that is trying to quickly visualise these types of data (as part of their exploratory data analysis, say). The `ggmatplot` R-package seeks to provide 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` after data wrangling using `dplyr` and `tidyr` (Wickham et al., 2019). We will use the example data in Tables 1 and 3, which are stored in the objects `wide_df` and `SnowGR`, respectively.

Example 1

The code below constructs a line plot (superimposed with a point) of the various types (food, service and ambience) of ratings, contained in columns 2 to 4 of `wide_df`, against the overall restaurant rating in column 5 of `wide_df` as shown in Figure 1.

In example 1, should X lab be restaurant or rating? The one made by ggmatplot is different from the one made by ggplot. Currently, the plot does not show a clear relationship between various types and the overall rating. Is it better to compare the ratings among restaurants? – commented by XL

```
library(ggmatplot)
ggmatplot(x = wide_df[, 2:4], y = wide_df[, 5], plot_type = "both",
          xlab = "Rating", ylab = "Overall rating", legend_title = "Type")
```

In contrast to the above, using `ggplot2` alone, the data must be wrangled to a long form first before plotting, as exemplified in the code below, in order to obtain the same result as Figure 1. This adds a small, but noticeable, friction to the workflow for the practitioner that is looking to promptly explore their data.

```
library(ggplot2)
library(tidyr) # or library(tidyverse)
wide_df %>%
  select(contains("rating")) %>%
  pivot_longer(-overall_rating,
               names_to = "rating_type",
               values_to = "rating") %>%
  ggplot(aes(rating, overall_rating, color = rating_type)) +
  geom_point(aes(shape = rating_type)) +
  geom_line(aes(group = rating_type, linetype = rating_type)) +
  labs(x = "Restaurant", y = "Rating",
       color = "Type", linetype = "Type", shape = "Type")
```

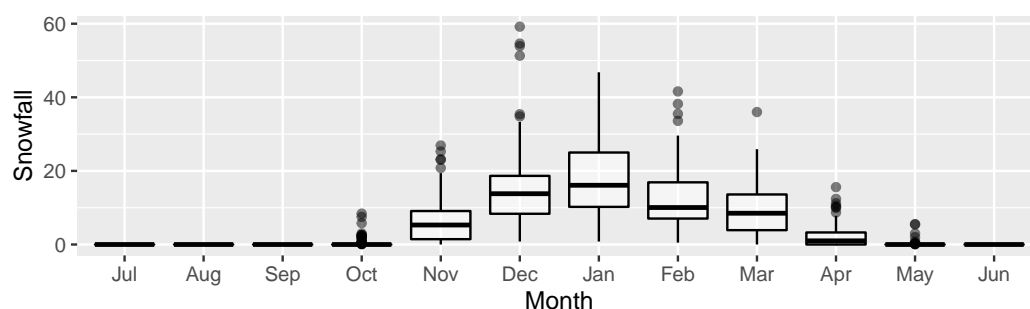


Figure 2: The distribution of the amount of snowfall at Grand Rapids, Michigan, across months from 1893-2011.

Example 2

The example code draws the boxplot of each column of amount of snowfall across months in the `SnowGR` data as shown in Figure 2. As the resulting object is a `ggplot` object, the user can leverage the `ggplot` functions to modify the output (e.g. removal of the legend). CAN WE `fct_inorder` SO THAT THE MONTHS APPEAR IN ORDER OF THEIR COLUMN NAMES? EMI HAS ALREADY PICKED THIS UP AS AN ISSUE...

Now the color and fill has been removed for boxplot. So by default, there is no legend. We may need to change the demonstration of Example 2. – commented by XL Another question is that only the github version has such setting. Do I need to submit the new version to CRAN? – commented by XL

```
library(ggmatplot)
ggmatplot(x = SnowGR[, 3:14], plot_type = "boxplot",
          xlab = "Month", ylab = "Snowfall") +
  guides(color = "none", fill = "none")
```

The equivalent code for the above to produce Figure 2 without using `ggmatplot` is given below. Again, we observe a slight but non-negligible friction in putting the data in the right format prior to plotting. The original wide data format like those shown in Table 3 is common in the environmental sciences among other disciplines, and thus an analyst who has to repeat these tasks can benefit from a quick approach as `ggmatplot` offers.

```
library(ggplot2)
library(tidyr) # or library(tidyverse)
SnowGR %>%
  pivot_longer(Jul:Jun,
               names_to = "Month",
               values_to = "Snowfall") %>%
  ggplot(aes(Month, Snowfall)) +
  geom_boxplot(aes(color = Month, fill = Month), alpha = 0.5) +
  guides(color = "none", fill = "none")
```

WOULD IT BENEFIT FROM ALSO SHOWING A PLOT THAT BASE MATPLOTT CAN NOT? THE ABOVE CAN BE DONE BY APPLY `boxplot`, AND HENCE WHY I WROTE VIOLIN PLOT AT THE BEGINNING AS IT IS NOT AVAILABLE AT LEAST IN CORE R?

The other option is the density plot which can not be done by `matplot`—commented by XL. But since there are too many months, it might not be a good idea. – commented by XL

Discussion

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

Acknowledgements

FKCH was supported by an Australian Research Council Discovery Fellowship DE200100435.

References

- Pruim, R., Kaplan, D., & Horton, N. (2021). *mosaicData: Project MOSAIC data sets*. Retrieved from <https://CRAN.R-project.org/package=mosaicData>
- 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, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10), 1–23. doi:[10.18637/jss.v059.i10](https://doi.org/10.18637/jss.v059.i10)
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <https://ggplot2.tidyverse.org>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemond, G., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. doi:[10.21105/joss.01686](https://doi.org/10.21105/joss.01686)
- Wilkinson, L. (2005). *The grammar of graphics*. Springer.