The Journal of Open Source Software

gg1d: visualizations for exploratory data analysis using tiled one-dimensional graphics and parallel coordinate plots

Sam El-Kamand [1], **Julian M. W. Quinn** [1], **and Mark J. Cowley** [1,2¶]

1 Children's Cancer Institute, Australia **2** School of Clinical Medicine, UNSW Medicine & Health, Australia ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Exploratory data analysis (EDA) involves examining relationships between both categorical and quantitative features. The gg1d R package streamlines EDA by providing two turnkey approaches to visualising n-dimensional data which can graphically reveal correlative or associative relationships between 2 or more features. For small datasets (n<1000), gg1d can represent all dataset features as distinct, vertically aligned bar or tile plots, with plot types auto-selected based on whether variables are categorical or numeric (Figure 1). Since datasets with thousands of observations are often challenging to visualise in tiled layouts, gg1d also produces interactive parallel coordinate plots (PCPs) better suited for examining large datasets (Figure 2). gg1d reduces the amount of code and time required to detect multi-feature relationships that may otherwise require statistical modelling or thorough manual review to identify (Figure 3, Figure 4).

To make gg1d visualisations accessible to a wider audience, we also developed EDA, a web app that enables non-programmers to explore and interpret data patterns interactively (Figure 5). EDA is available at https://selkamand.github.io/EDA/.

The Journal of Open Source Software

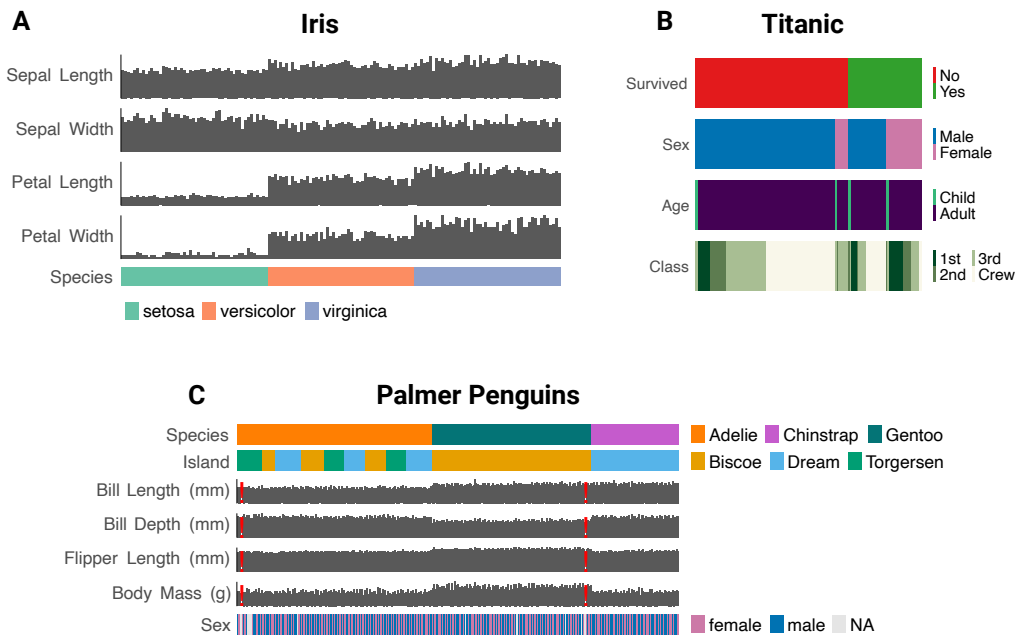

Figure 1: **gg1d** visualizations of common datasets revealing: a) Petals of the *setosa* species of iris are drastically smaller than other iris species; b) The majority of individuals who perished during the Titanic disaster were adult males; c) *Gentoo* penguins from Biscoe Island have shallower bill depths than *Chinstrap* or *Adelie* penguins, despite their increased body mass. Exclamation marks indicate missing values.

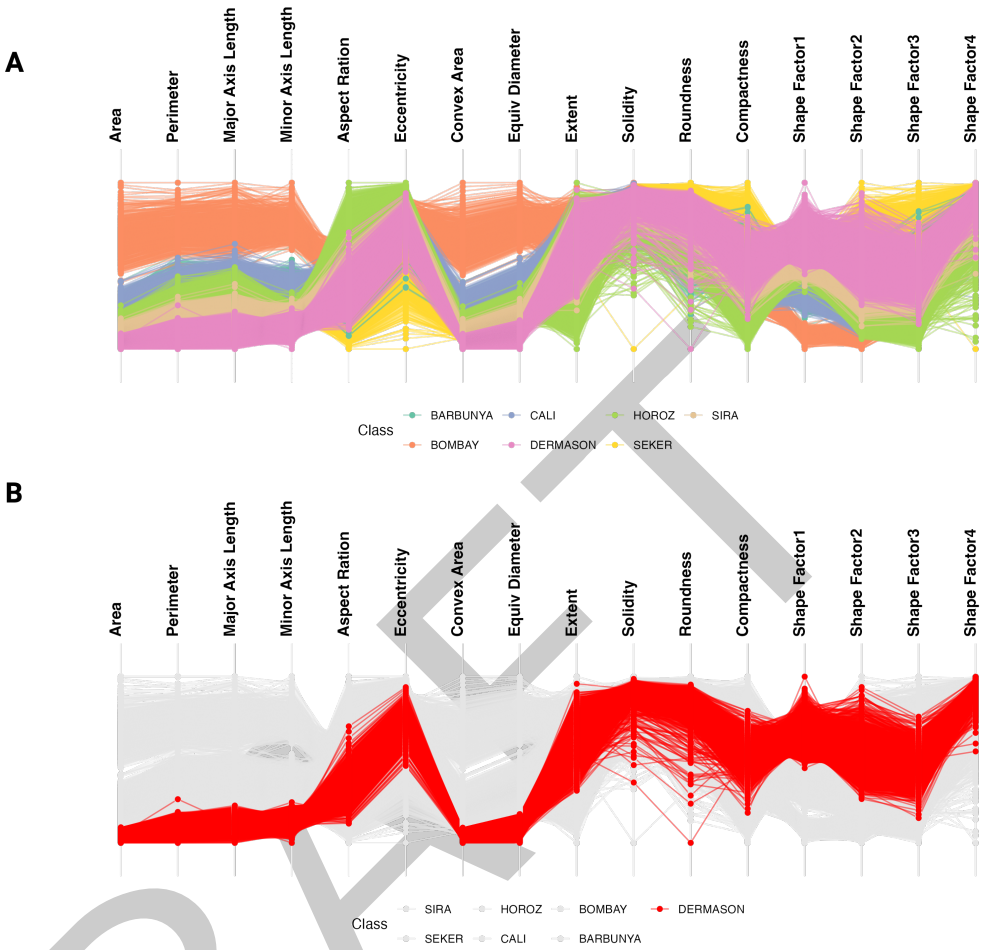

Figure 2: **gg1d** parallel coordinate plots of the dry beans imaging dataset (**?**). a) Visualizing 16 morphological features of 13,611 grains from common dry bean species reveals clear correlations amongst size-related attributes (Area, Perimeter and Axis Length). Bombay beans were the largest, most convex variety; b) Highlighting a single subclass simplifies both comparison against the full cohort and identification of within-class outliers. For example, Dermason beans (red) are smaller in size than other varieties. One Dermason bean grain had unusually low roundness, highly atypical for this variety

Statement of Need

The R ecosystem already includes popular EDA packages such as **skimr**, which textually summarizes completeness and descriptive statistics for individual features (1-dimensional), and **GGally**, which graphically describes pairwise feature correlations (2-dimensional) or multi-feature relationships through PCPs (n-dimensional). **gg1d** enhances this ecosystem by providing interactive versions of standard n-dimensional visualisations like PCPs and introducing tiled one-dimensional visualisations that more effectively show missingness and complex categorical relationships in smaller datasets. Together, these visualisations provide key advantages over other EDA packages, most notably their ability to reveal a greater variety of complex multidimensional patterns (Figure 3, Figure 4).

Feature	gg1d	Complex Heatmap	Data Explorer	skimr	GGally	ggpcp
Automatic Plot Generation	✓	✗	✓	✓	✓	✗
Automatic plot selection by variable type	✓	✗	✓	✗	✓	✗
Interactive Visualisations	✓	✗	✗	✗	✗	✗
Supports cross-linking with other datasets	✓	✗	✗	✗	✗	✗
Composable with Patchwork	✓	✓	✗	✗	✓	✓
Supports parallel coordinate plots	✓	✗	✗	✗	✓	✓
Describes features contribution to total variance (PCA)	✗	✗	✓	✗	✗	✗
Generates Publication Quality Figures	✓	✓	✗	✗	✓	✓
Graphical User Interface is available	✓	✓	✗	✗	✗	✗
Reveals missingness dependent on multiple features	✓	✗	✗	✗	✗	✗

Figure 3: Comparison of R packages that create visualisations commonly used for exploratory data analysis, including ComplexHeatmap (Gu, 2022), Data Explorer (Cui, 2024), skimr (Waring et al., 2022), GGally [Schloerke:2024] and ggpcp (Susan VanderPlas & Hofmann, 2023). Due to documented reproducibility issues, ggpcp features could not be verified first-hand.

The benefits of **gg1d** are exemplified when visualizing the artificial Lazy Birdwatcher dataset, which records magpie observations by two birdwatchers (Figure 4). One birdwatcher does not work on weekends, creating a missing data pattern dependent on both birdwatcher and day of the week. This multidimensional pattern becomes immediately apparent from the **gg1d** tile plots, whereas it is difficult to detect using one-dimensional EDA tools like **skimr**, two-dimensional tools like **ggpairs** from the **GGally** package. Despite being n-dimensional, all PCP plot implementations in R also fail to uncover this trend due to either exclusion of missing data or inability to effectively represent relationships between categorical features (data not shown).

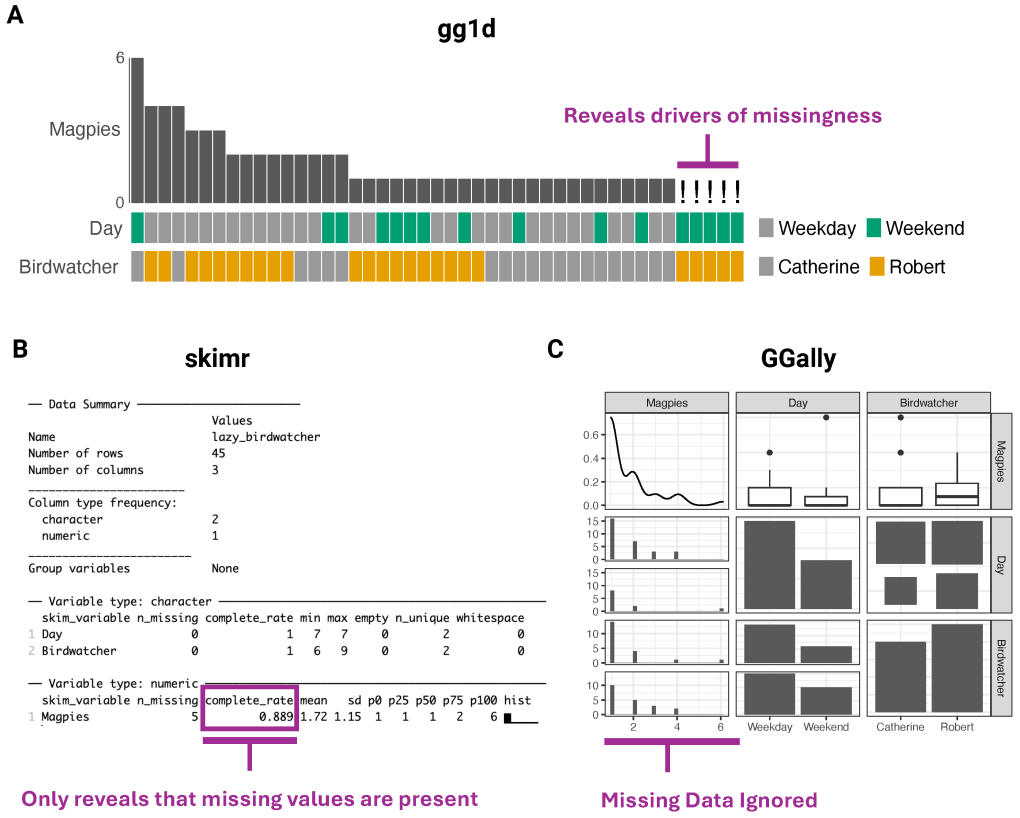

Figure 4: Visualisation of the Lazy Birdwatcher dataset using the **gg1d** package reveals a pattern of missingness dependent on multiple variables, Birdwatcher and Day (A). This pattern is difficult to detect using one-dimensional EDA tools like **skimr** (B) or two-dimensional tools like ggpairs from the **GGally** package (C).

Despite the advancements provided by gg1d and other tools in the R ecosystem, a key limitation remains: accessibility for non-programmers, particularly when visualizing n-dimensional data. All existing R implementations lack graphical user interfaces (Figure 3). While shiny web apps offer a potential solution, they often require uploading datasets to external servers, raising privacy concerns. To address these limitations, we developed EDA, a web-assembly compiled client-side web app for secure, interactive data exploration (Figure 5). Operating entirely in the browser, EDA ensures data remains on the user's machine, increasing ease of use without compromising data privacy. EDA is available at https://selkamand.github.io/EDA/.

The Journal of Open Source Software

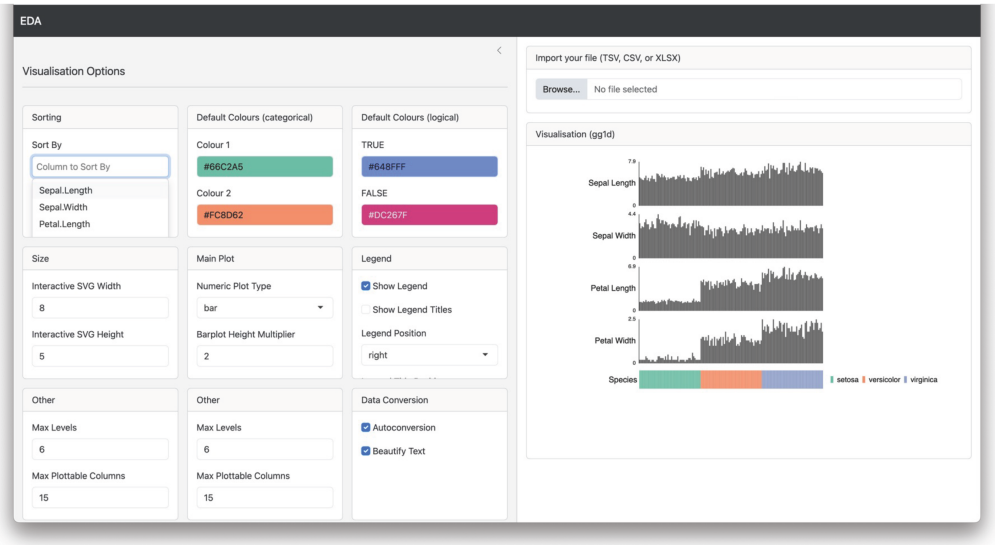

Figure 5: Screenshot of **EDA**, a web-app providing a graphical user interface for code-free generation of **gg1d** visualisations.

We developed gg1d and associated tools for the visualisation of clinical and multiomics data and anticipate it will prove valuable for any exploratory EDA activities.

Acknowledgements

We thank the developers of the packages integral to gg1d, especially David Gohel for ggiraph (Gohel & Skintzos, 2024), which enables its interactivity, and Thomas Lin Pedersen for patchwork (Pedersen, 2024) and ggplot2 maintenance. We also acknowledge Hadley Wickham and all contributors to ggplot2 (Wickham, 2016). The gg1d graphical user interface (EDA) was made possible thanks to creators and maintainers of shiny (Chang et al., 2024), shinylive (Schloerke et al., 2024) and webR (Stagg et al., 2023).

References

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). *Shiny: Web application framework for r*. https://doi.org/https://doi.org/10.32614/CRAN.package.shiny

Cui, B. (2024). *DataExplorer: Automate data exploration and treatment*. https://doi.org/https://doi.org/10.32614/CRAN.package.DataExplorer

Gohel, D., & Skintzos, P. (2024). *Ggiraph: Make 'ggplot2' graphics interactive*. https://doi.org/10.32614/CRAN.package.ggiraph

Gu, Z. (2022). Complex heatmap visualization. *iMeta*, *1*(3), e43. https://doi.org/10.1002/imt2.43

Pedersen, T. L. (2024). *Patchwork: The composer of plots*. https://doi.org/10.32614/cran.package.patchwork

Schloerke, B., Chang, W., Stagg, G., & Aden-Buie, G. (2024). *Shinylive: Run 'shiny' applications in the browser*. https://CRAN.R-project.org/package=shinylive

Stagg, G. W., Lionel, H., & Others. (2023). *webR: The statistical language r compiled to WebAssembly via emscripten*. https://docs.r-wasm.org/webr/latest/

Susan VanderPlas, A. U., Yawei Ge, & Hofmann, H. (2023). Penguins go parallel: A grammar of graphics framework for generalized parallel coordinate plots. *Journal of Computational and Graphical Statistics*, *32*(4), 1572–1587. https://doi.org/10.1080/10618600.2023.2195462

Waring, E., Quinn, M., McNamara, A., Arino de la Rubia, E., Zhu, H., & Ellis, S. (2022). *Skimr: Compact and flexible summaries of data*. https://doi.org/https://doi.org/10.32614/CRAN.package.skimr

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://doi.org/10.32614/CRAN.package.ggplot2