gg1d: Visualizing Multidimensional Patterns in R

17 June 2024

# Summary

Exploratory data analysis (EDA) begins by assessing dataset completeness and evaluating interrelationships among both categorical and quantitative features. Visualizing multiple variables simultaneously can reveal multidimensional patterns (Figure 1), including those driving missingness or explaining outliers (Figure 2). The gg1d R package stacks visual representations of each feature in a dataset on top of one another, facilitating rapid exploration of features covariance. Visualizations can be sorted by a feature of interest to draw out further insights (Figure 1B). The gg1d package is available on GitHub at <https://github.com/selkamand/gg1d>. A gallery of example gg1d visualizations, including those shown below, is available [here](https://selkamand.github.io/gg1d/articles/gallery.html#artificial-data).

Default gg1d output from common datasets revealing a) the Setosa species of iris has drastically smaller petals then other species, b) The vast majority of individuals that perished during the titanic were adult males and c) Gentoo penguins from Biscoe Island have shallower bill depths than Chinstrap or Adelie, despite their increased body mass. Exclamation mark symbols indicate missing values. 

Default gg1d output from common datasets revealing a) the Setosa species of iris has drastically smaller petals then other species, b) The vast majority of individuals that perished during the titanic were adult males and c) Gentoo penguins from Biscoe Island have shallower bill depths than Chinstrap or Adelie, despite their increased body mass. Exclamation mark symbols indicate missing values.

# Statement of Need

The R ecosystem includes popular EDA packages including **skimr**, which textually summarizes completeness and statistics for individual features (1-dimensional), and **GGally**, which graphically describes pairwise feature correlations (2-dimensional). gg1d is simply an n-dimensional generalisation that can reveal multidimensional patterns.

The benefits of gg1d are exemplified when visualizing the (artificial) **Lazy Birdwatcher** dataset, which records magpie observations by two birdwatchers (Figure 2). One birdwatcher does not work on weekends, creating a missing data pattern dependent on both birdwatcher and day of the week. This multidimensional pattern is immediately obvious in gg1d output but difficult to detect using one-dimensional EDA tools like skimr or two-dimensional tools like ggpairs from the **GGally** package.

gg1d visualization of the Lazy Birdwatcher dataset revealing a pattern of missingness dependent on multiple variables (birdwatcher and weekend). This pattern is difficult to detect using one-dimensional EDA tools like skimr or two-dimensional tools like ggpairs from the GGally package. 

gg1d visualization of the Lazy Birdwatcher dataset revealing a pattern of missingness dependent on multiple variables (birdwatcher and weekend). This pattern is difficult to detect using one-dimensional EDA tools like skimr or two-dimensional tools like ggpairs from the GGally package.

## Other applications

# Acknowledgements

We thank the developers of the packages integral to gg1d, especially David Gohel for ggiraph [@gohel: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].

# References