Make your paper figures professionally: Scientific data analysis and visualization with R

Vangelis Theodorakis, Fatemeh Behjati, Julien Gagneur, Marcel Schulz

18 October, 2020

Package

BiocStyle 2.16.1

Contents

1 Setup

```
library(ggplot2)
library(data.table)
library(magrittr)   # Needed for %>% operator
library(tidyr)
```

2 Choosing the appropriate visualization method

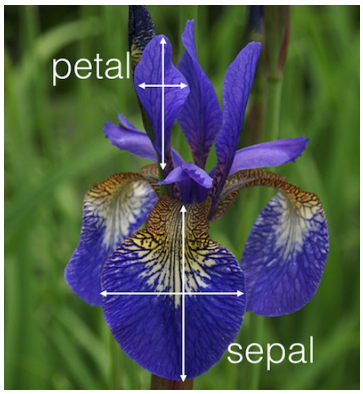
Match each chart type with the relationship it shows best.

1. shows distribution and quantiles, especially useful when comparing distributions.
2. highlights individual values, supports comparison, can show rankings or deviations categories and totals
3. shows overall changes and patterns, usually over time
4. shows relationship between two quantitative variables.

Options: bar chart, line chart, scatterplot, boxplot

3 Getting to know your dataset

`Iris` is a classical and widely used dataset in machine learning literature. It was first introduced by R.A. Fisher in his 1936 paper. Load the *iris* data into your R environment. What is the dimension of the dataset? What kind of data type does each column has? How many Species does it contain?



4 Data exploration

How are the lengths and widths of sepals and petals distributed? How would you visualize them? `Hint: tidy the data set and facet_wrap()`.

5 Misleading plots

1) Visualize the lengths and widths of the sepals and petals from the iris data with boxplots.
2) Add jitter (`geom_jitter()`) to visualize all points. Discuss: in this case, why is it not good to visualize the data with boxplots?
3) Alternatives to boxplot are violin plots (`geom_violin()`) and beanplots (`geom_beeswarm()` from library `ggbeeswarm`. Install it with `install.packages("ggbeeswarm")`). Apply both options to the same data.
4) Which pattern shows up when moving from boxplot to violin/bean plot? Give possible explanations for it. How could you prove your theories graphically?

6 Enchancing plots

Below is a graph taken from a published paper. Read the figure legend.

1) Discuss good and bad graphical properties of the plot. Make suggestions on how to improve it.
2) Implement a better visualization. As the original data is not available, we use the data simulated with the code below (also uploaded to Moodle).

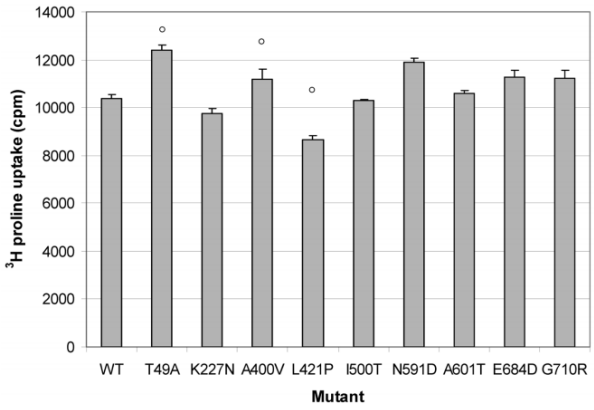


Figure 2. Maximal ^{3}H proline uptake of wildtype (WT) and all tested mutants. The maximum in uptake was measured in the presence of 3 µM cold L-proline. Data are expressed as means ± standard deviation (SD) obtained from triplicate samples. Mutants with a circle were tested in a second independent experiment.
doi:10.1371/journal.pone.0068645.g002

```
# simulate data
dt <- data.table(pro_uptake = c(
  rnorm(3, 10100, 300), rnorm(4, 12100, 300), rnorm(3, 9850, 300),
  rnorm(4, 11100, 300), rnorm(4,8300, 300), rnorm(3,10050, 300),
  rnorm(3, 12000, 300), rnorm(3, 10020, 300), rnorm(3, 10080, 300),
  rnorm(3, 10070, 300)),
  mutants = c(rep('WT',3), rep('T49A',4), rep('K227N',3), rep('A400V',4),
              rep('L421P',4), rep('I500T',3), rep('N591D',3), rep('A601T',3),
              rep('E684D',3), rep('G710R',3) )
)
```

6.1 Data analysis

Read the `titanic.csv` file. You can read description of the dataset from kaggle. Did age play a role in determining survival? Visualize this with a boxplot.

Now use facets to visualize whether age and gender combined were factors in survival. Do the same for age and passenger class.

Finally, visualize the interaction of age, gender and passenger class in determining survival.