

Data Visualization in R with **ggplot2**

The Choice of the Chart Type

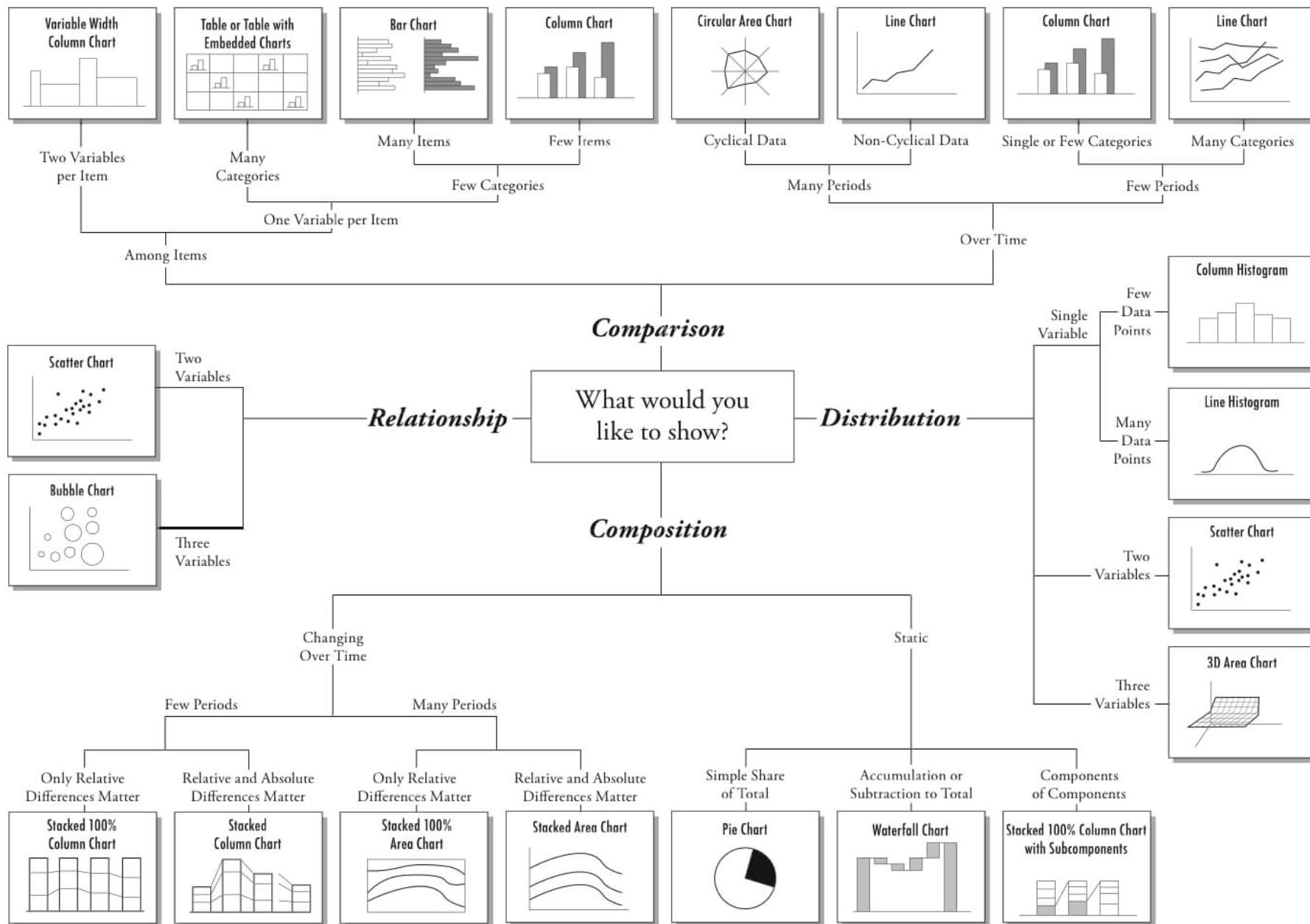
Cédric Scherer

Physalia Courses | March 2-6 2020

Photo by Richard Strozyński

Chart Suggestions—A Thought-Starter

www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com





from Data to Viz

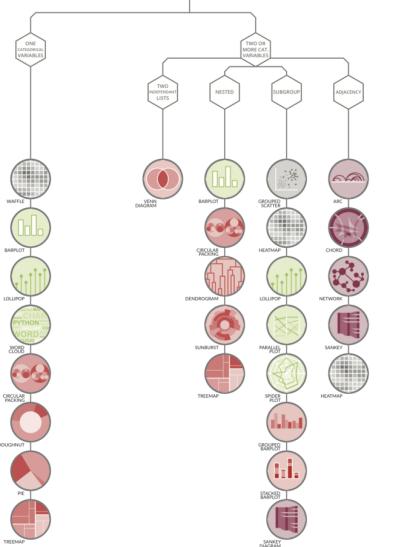
'From Data to Viz' is a classification of chart types based on input data format. It will help you find the perfect chart in three simple steps :

- 1** Identify what type of data you have.
- 2** Go to the corresponding decision tree and follow it down to a set of possible charts.
- 3** Choose the chart from the set that will suit your data and your needs best.

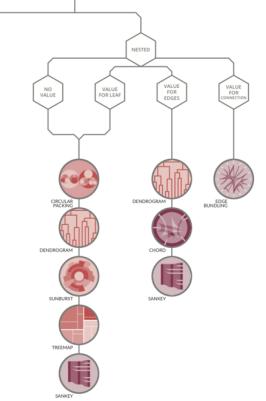
DataViz is a world with endless possibilities and this project does not claim to be exhaustive. However it should provide you with a good starting point. For an interactive version and much more, visit:

data-to-viz.com

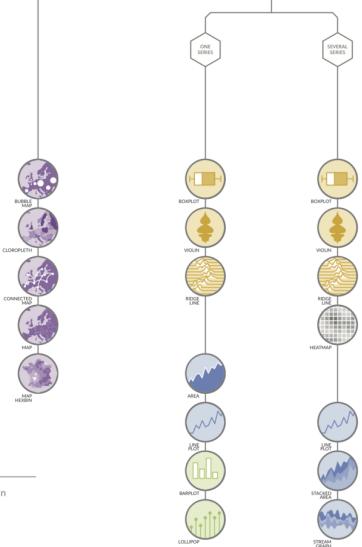
CATEGORIC



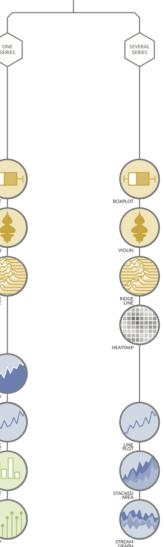
RELATIONAL



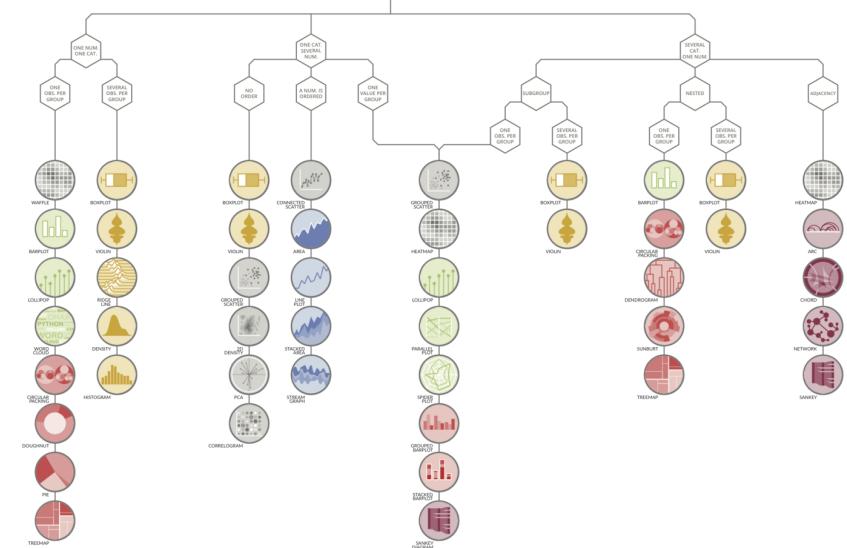
MAP



TIME SERIES

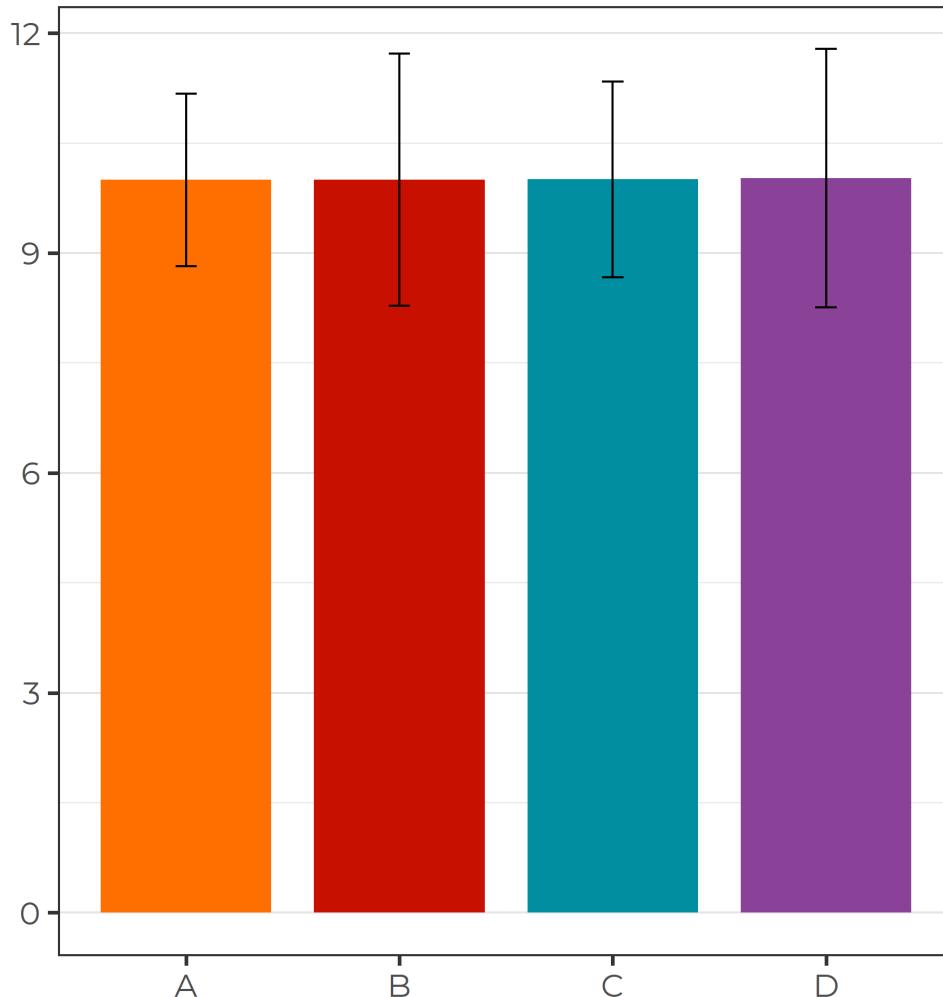


CATEGORIC AND NUMERIC



Barplot

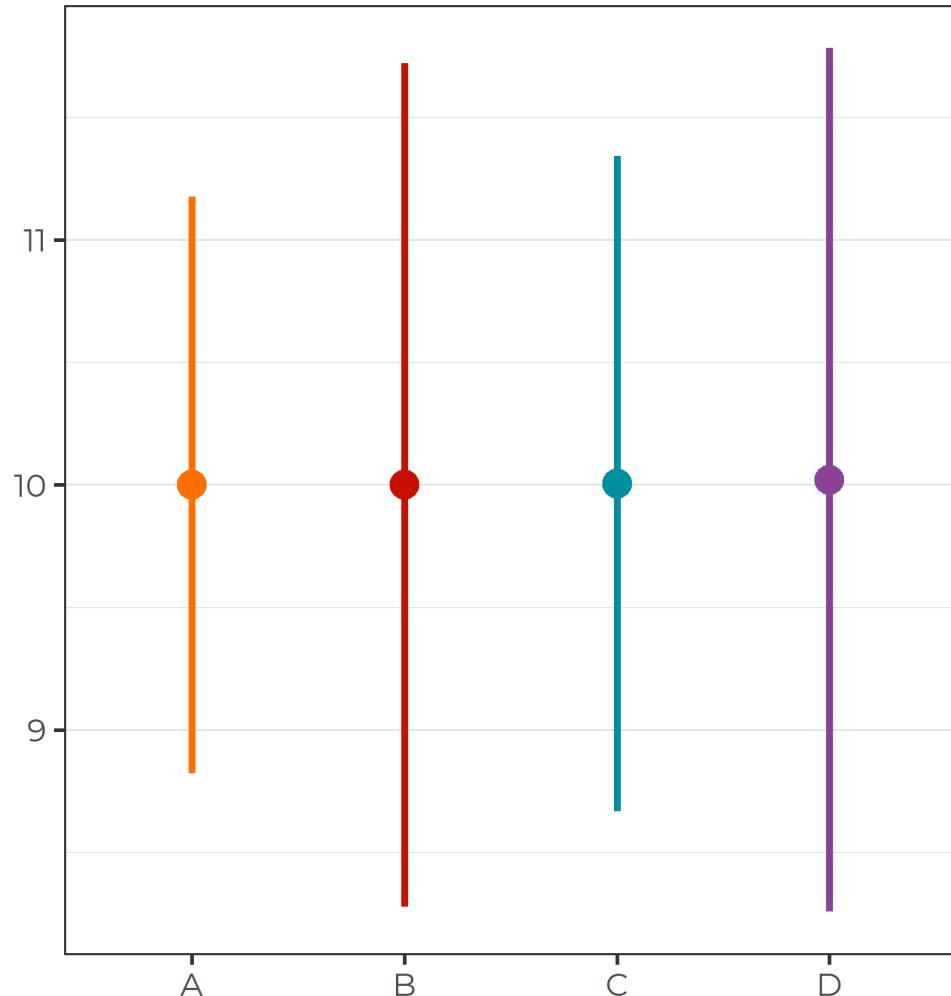
mean \pm SD



- Boring
- Not accurate
- Lots of useless space

Error Plot

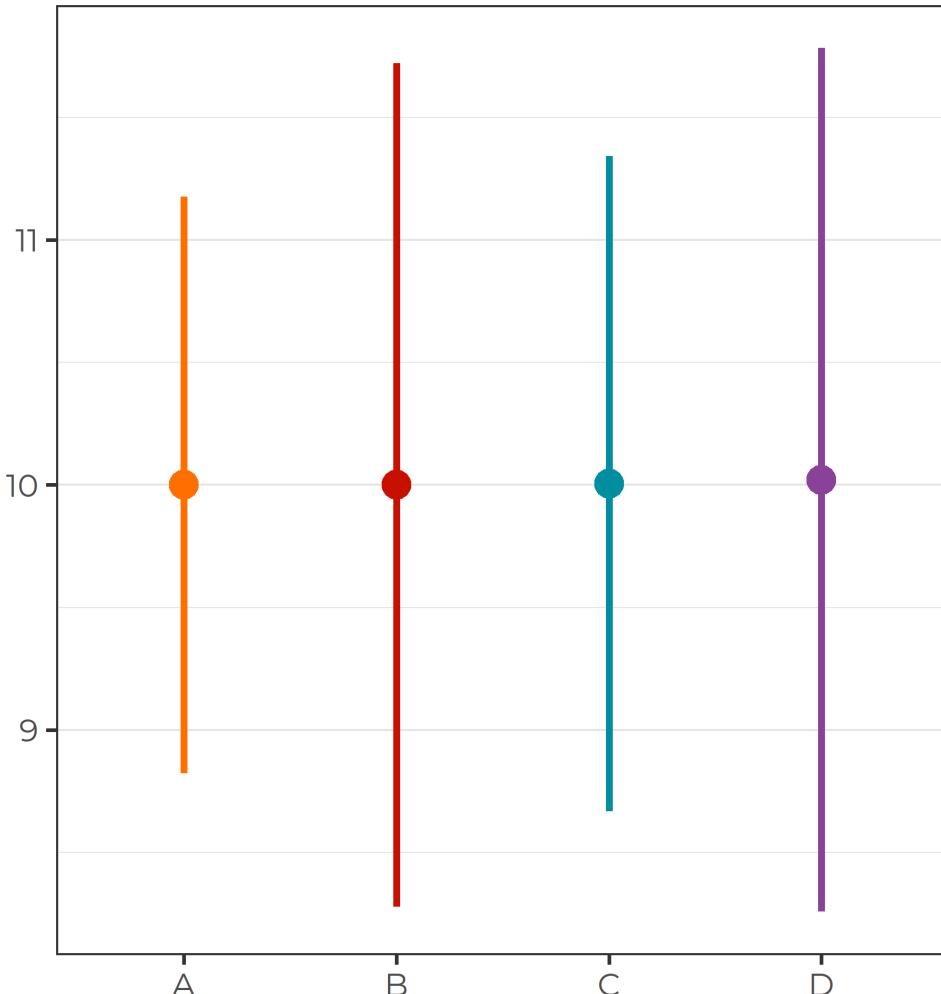
mean \pm SD



Visualization by Cédric Scherer

Error Plot

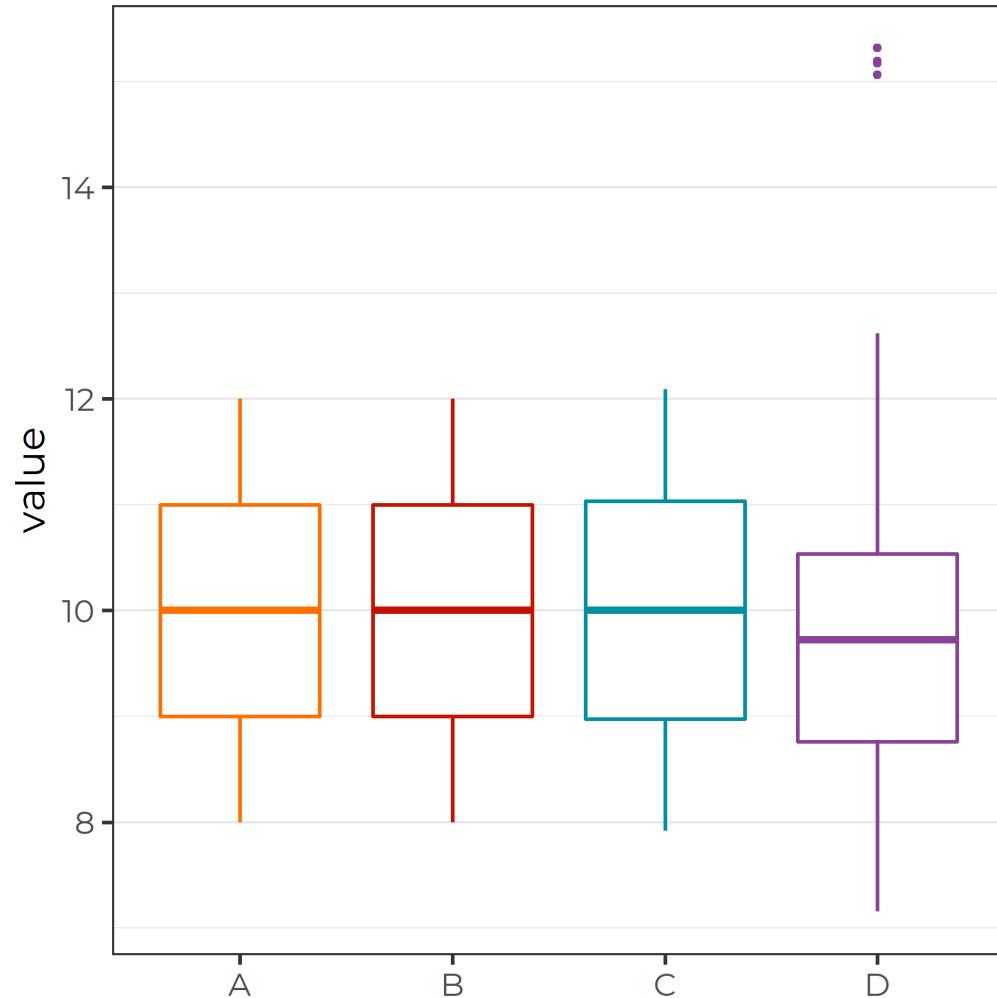
mean \pm SD



- Good: More accurate than a bar plot
- Good: Better data-ink ratio
- But: unclear what's shown

Box and Whiskers Plot

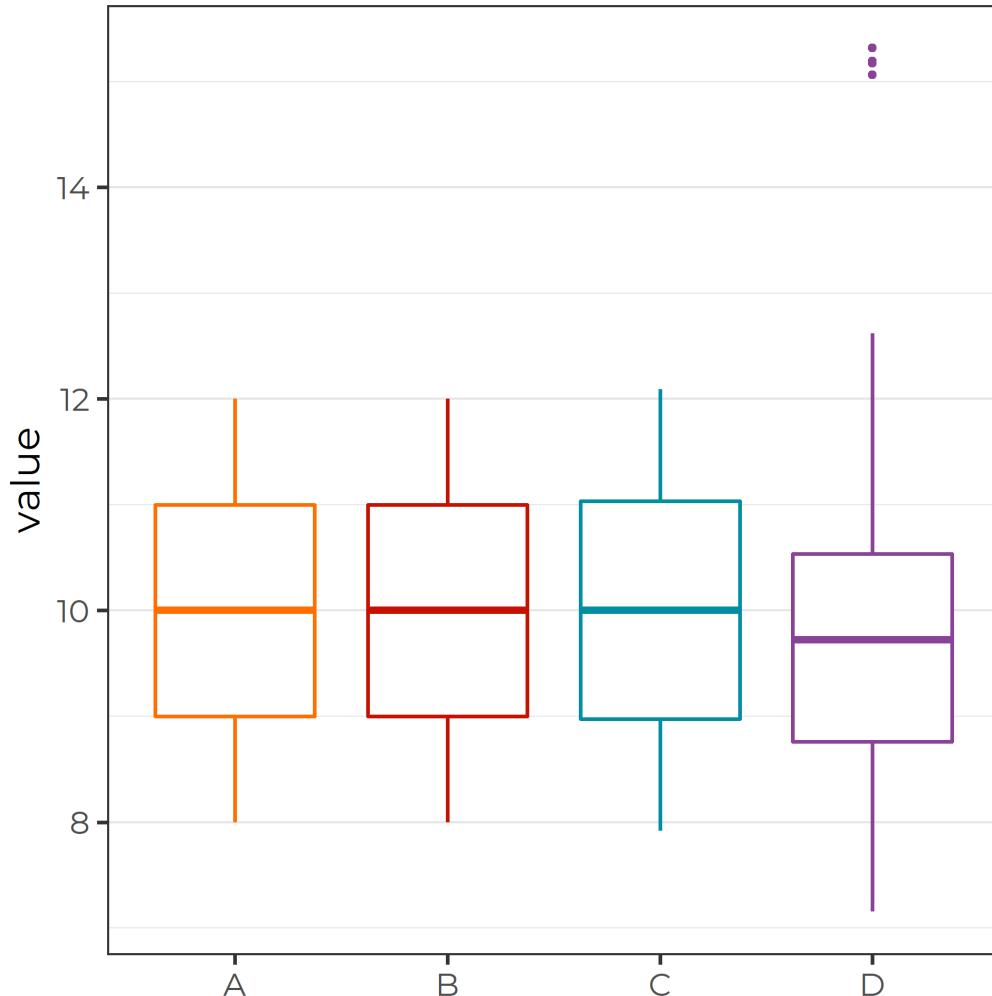
median, inter-quartile-range (IQR) and outliers



Visualization by Cédric Scherer

Box and Whiskers Plot

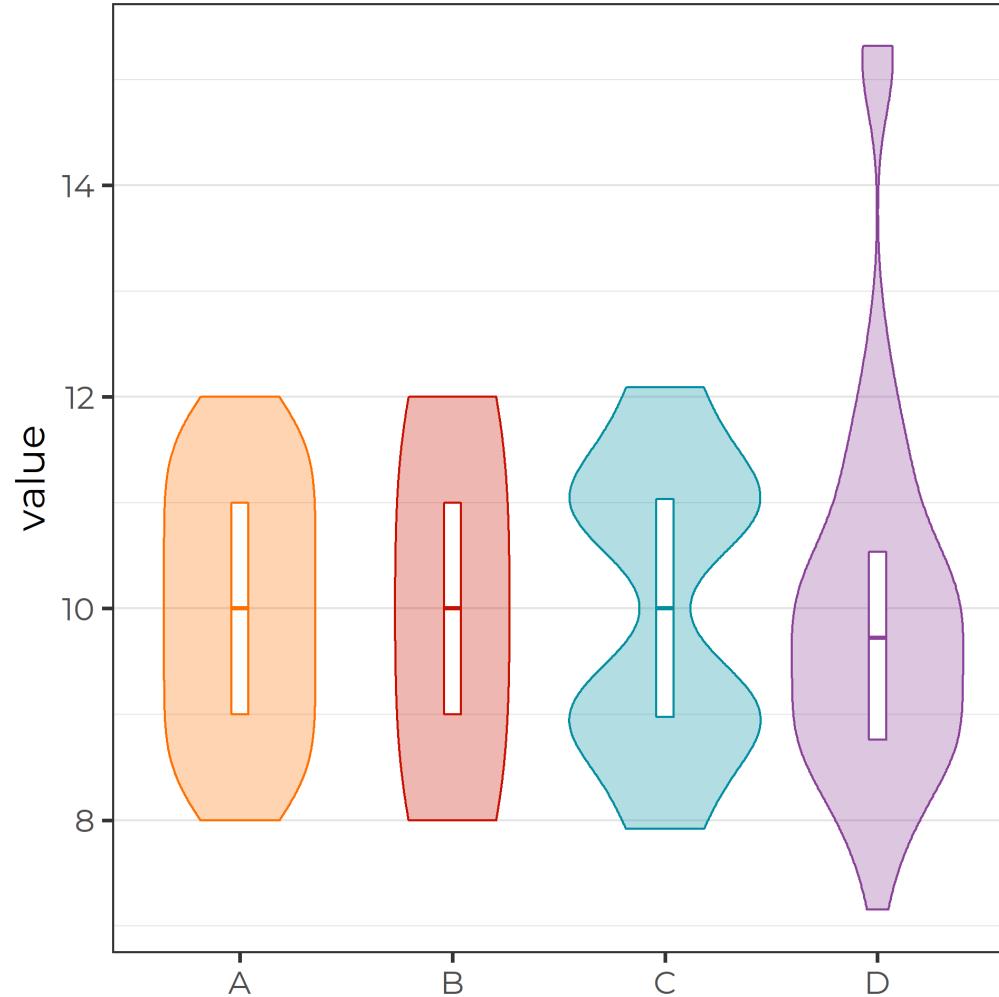
median, inter-quartile-range (IQR) and outliers



- Good: shows important summary stats
- Looks "scientifically"
- But: no info on sample size
- But: hard to grasp for broad audience

Violin Plot

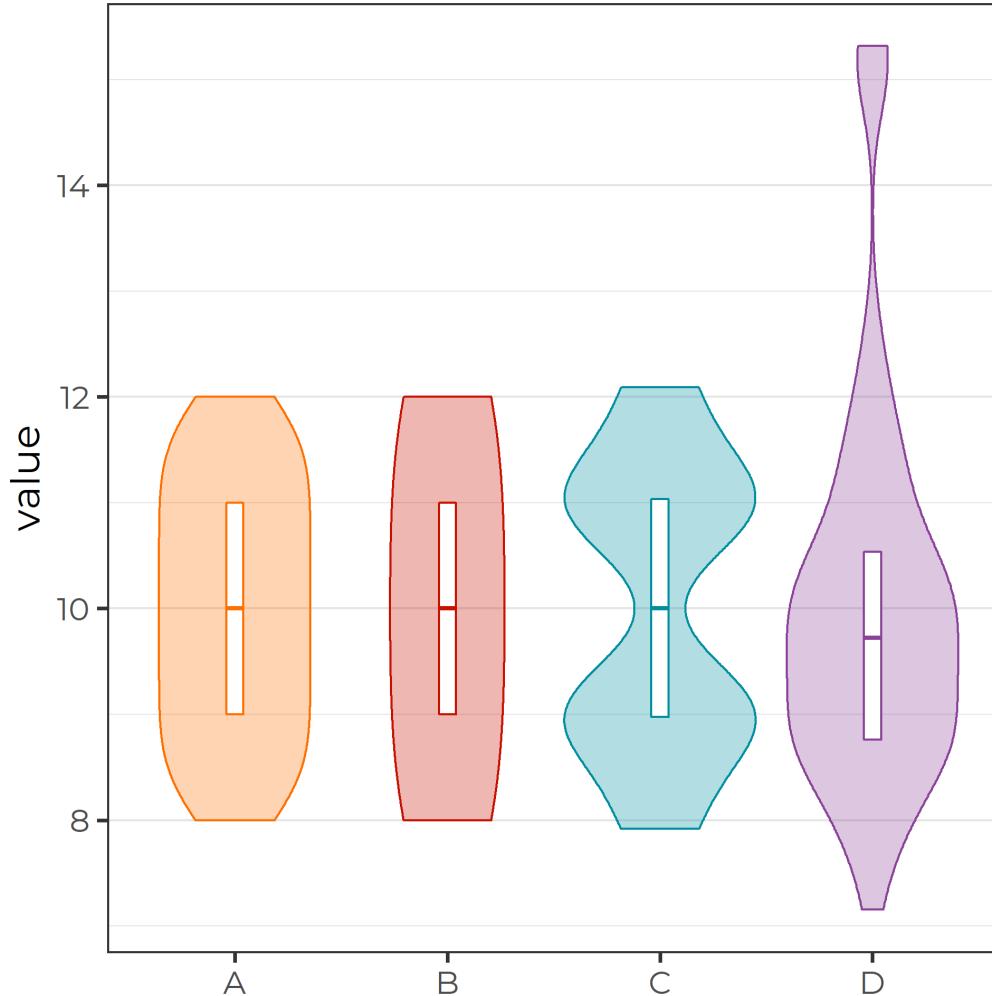
distribution, median and IQR



Visualization by Cédric Scherer

Violin Plot

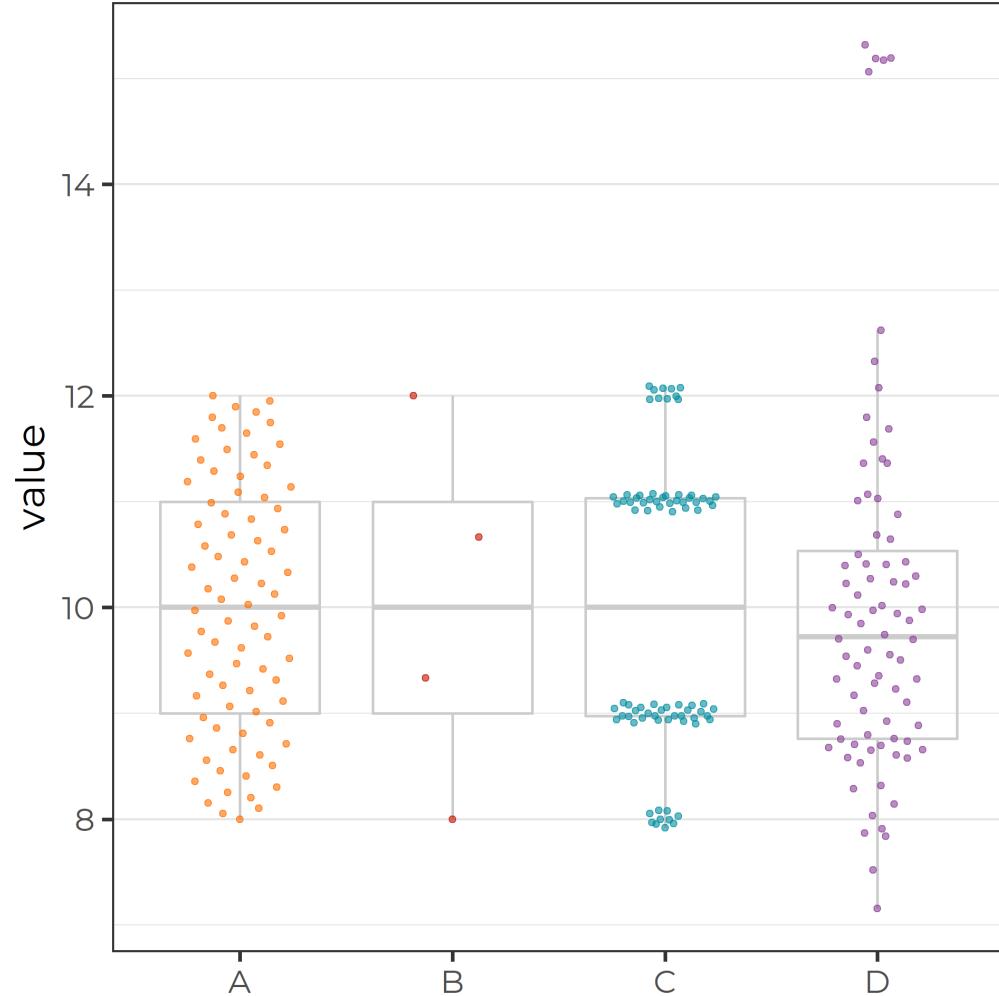
distribution, median and IQR



- Good: shows distribution + sample size
- But: sample size hard to estimate

Jitter or Sina Plot

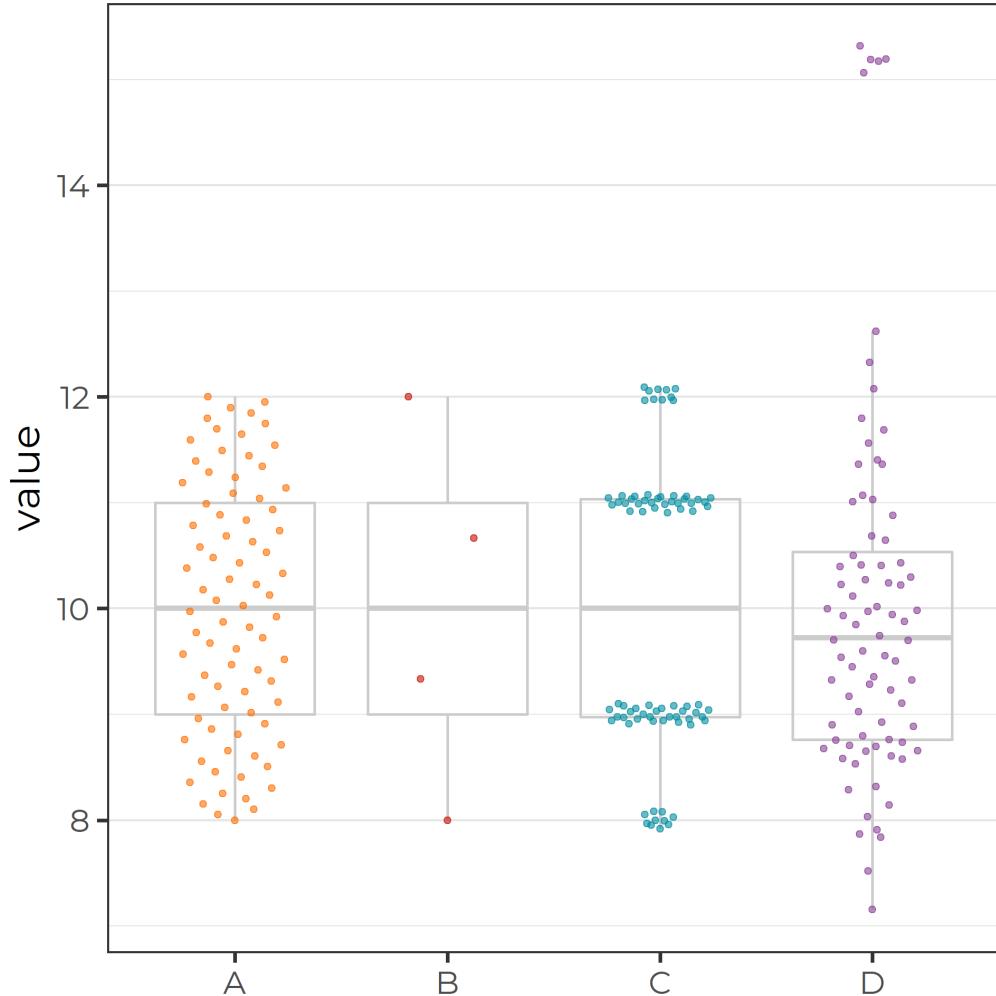
raw data (jittered)



Visualization by Cédric Scherer

Jitter or Sina Plot

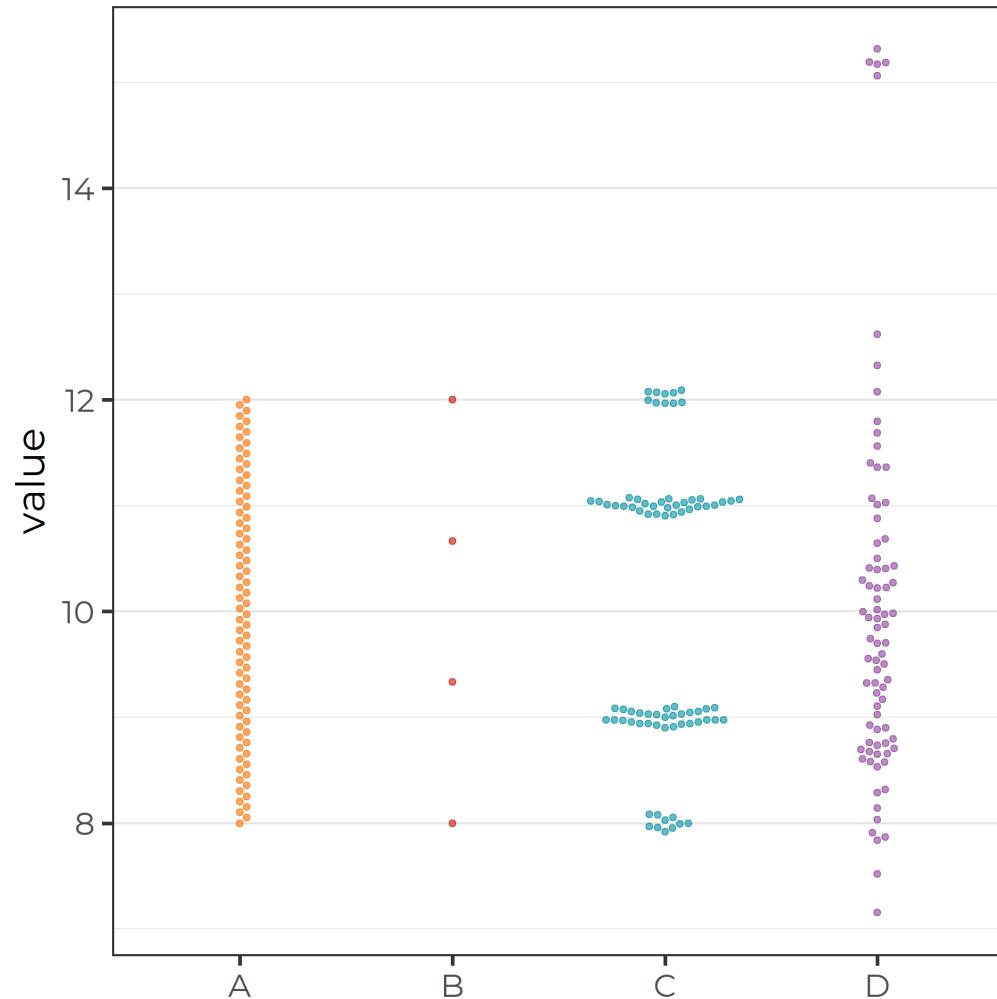
raw data (jittered)



- Good: shows sample size + distribution
- But: difficult with large sample sizes

Beeswarm Plot

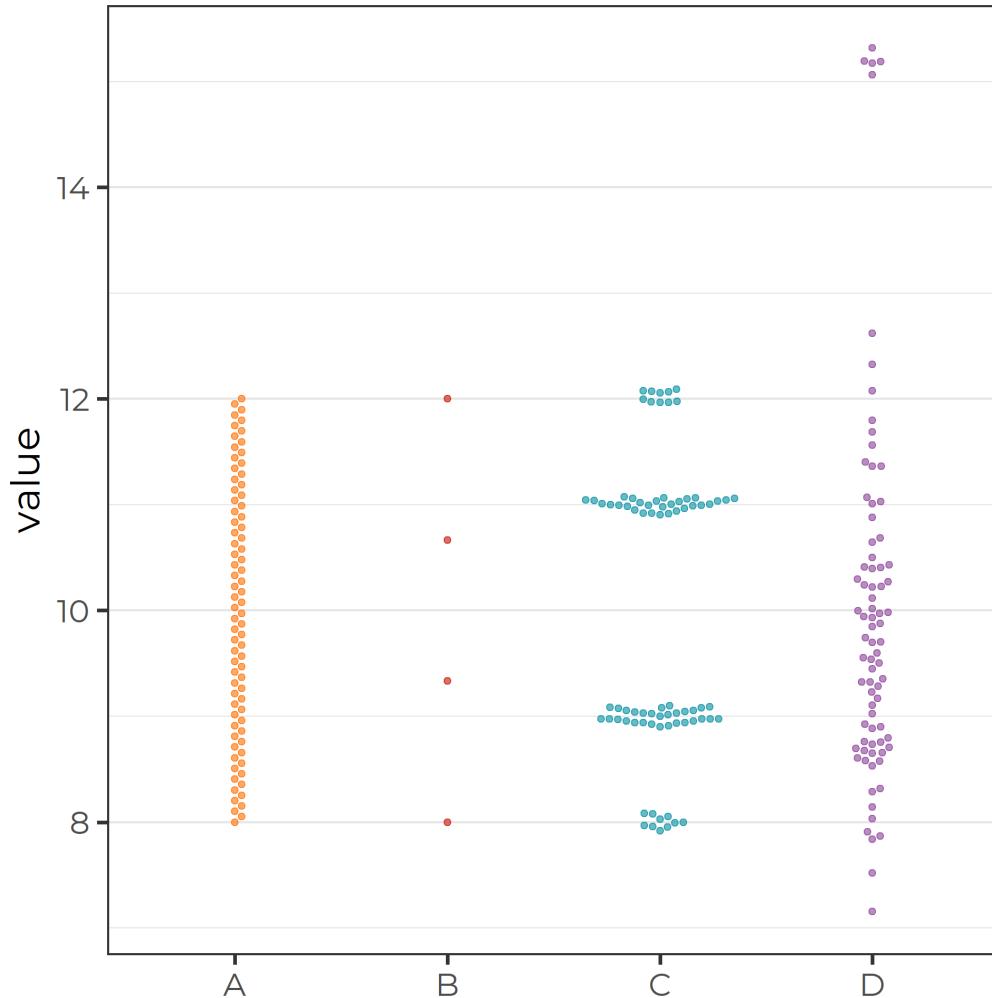
raw data without overlap



Visualization by Cédric Scherer

Beeswarm Plot

raw data without overlap

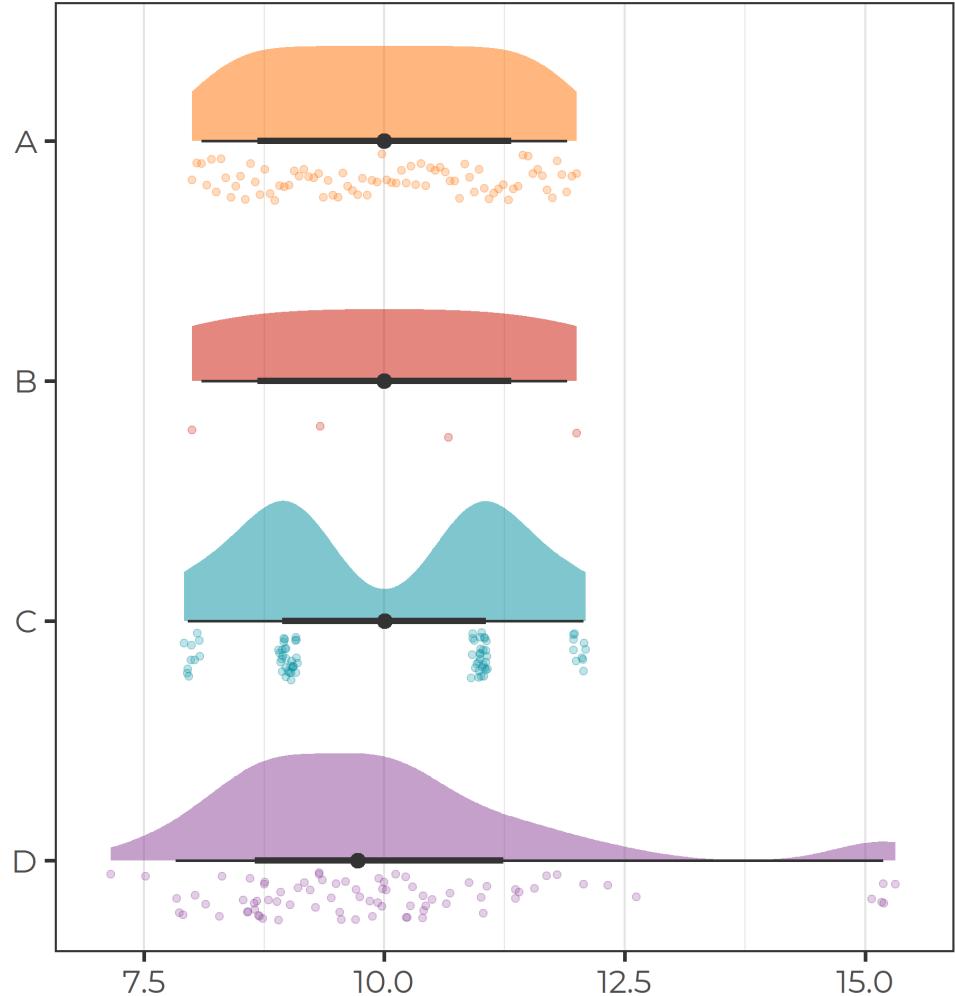


Visualization by Cédric Scherer

- Good: shows raw data + distribution
- Good: avoids overplotting
- But: difficult with large sample sizes

Raincloud Plot

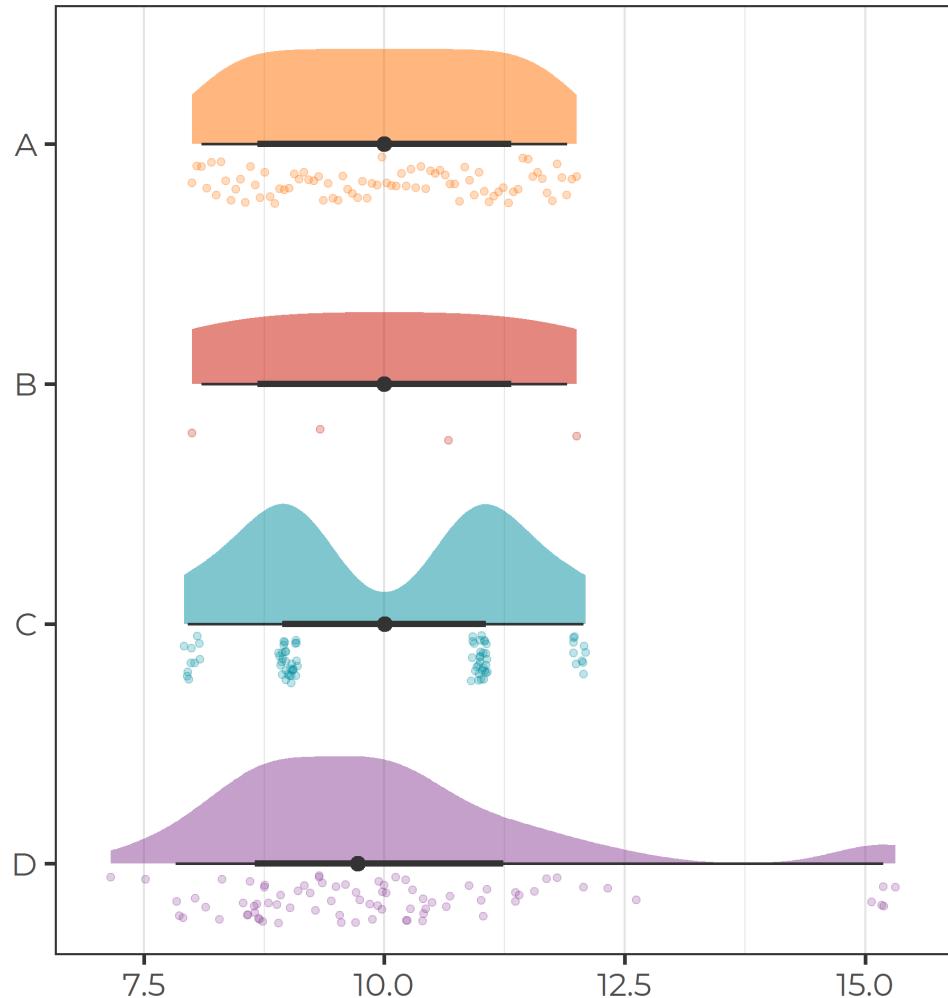
distribution, median, density and raw data



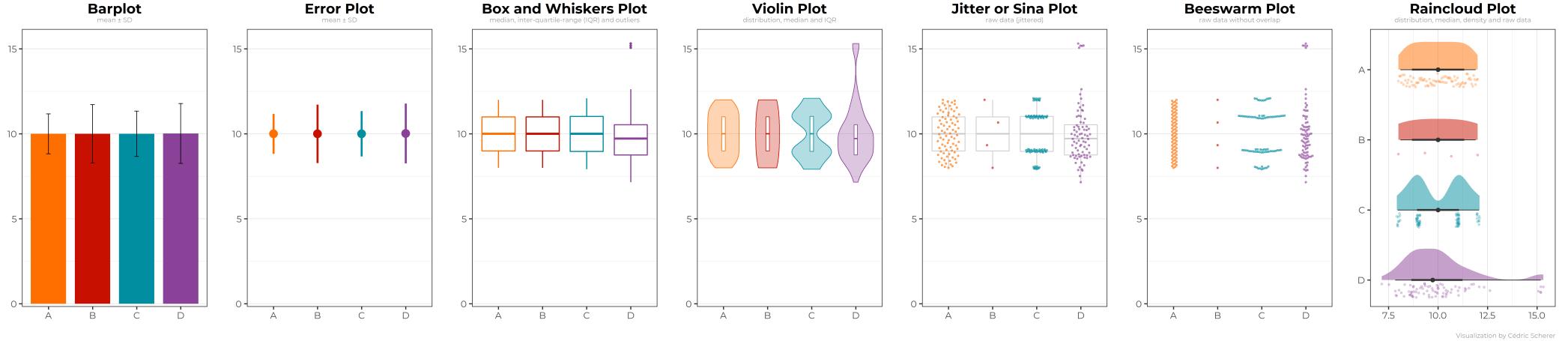
Visualization by Cédric Scherer

Raincloud Plot

distribution, median, density and raw data

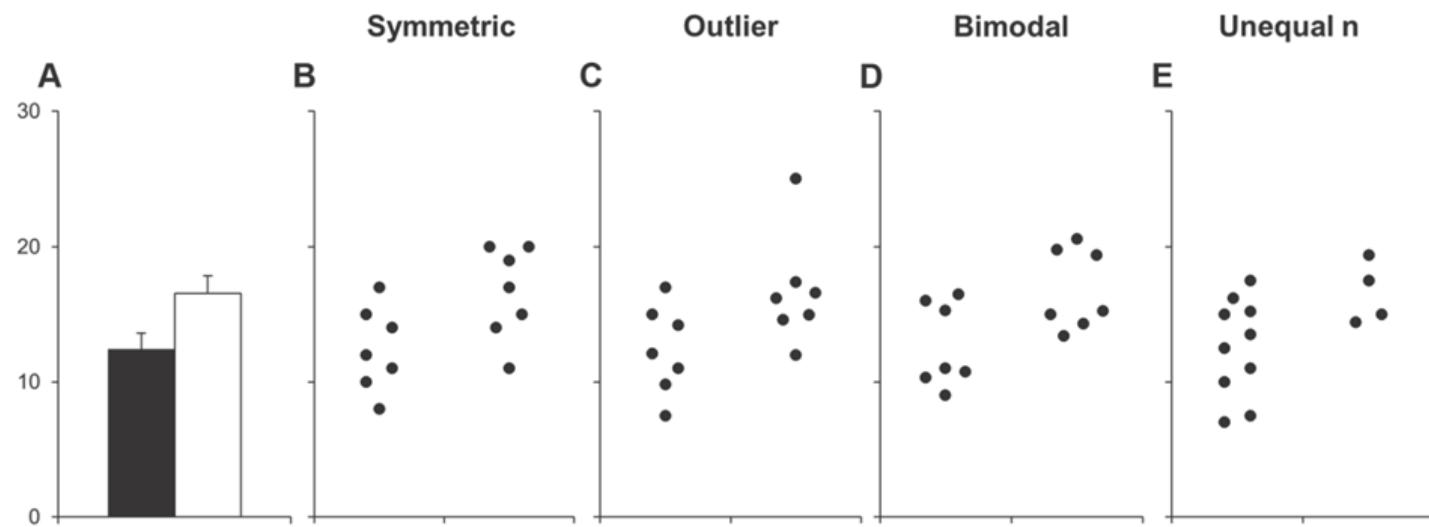


- Good: shows raw data + distribution
- Good: shows important summary stats
- But: needs more space



Visualization by Cédric Scherer

- Always check raw data and sample size
- Try several chart types
- Be open to combine chart types
- Choose chart type with your audience in mind



Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

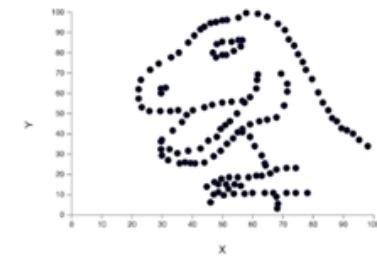
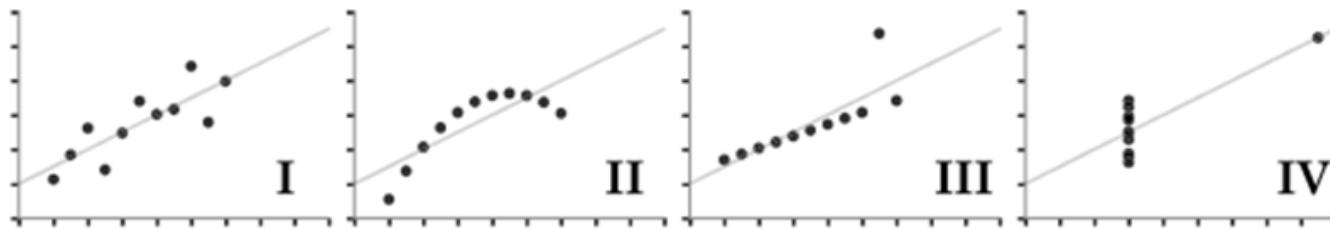


Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

Justin Matejka and George Fitzmaurice
Autodesk Research, Toronto Ontario Canada

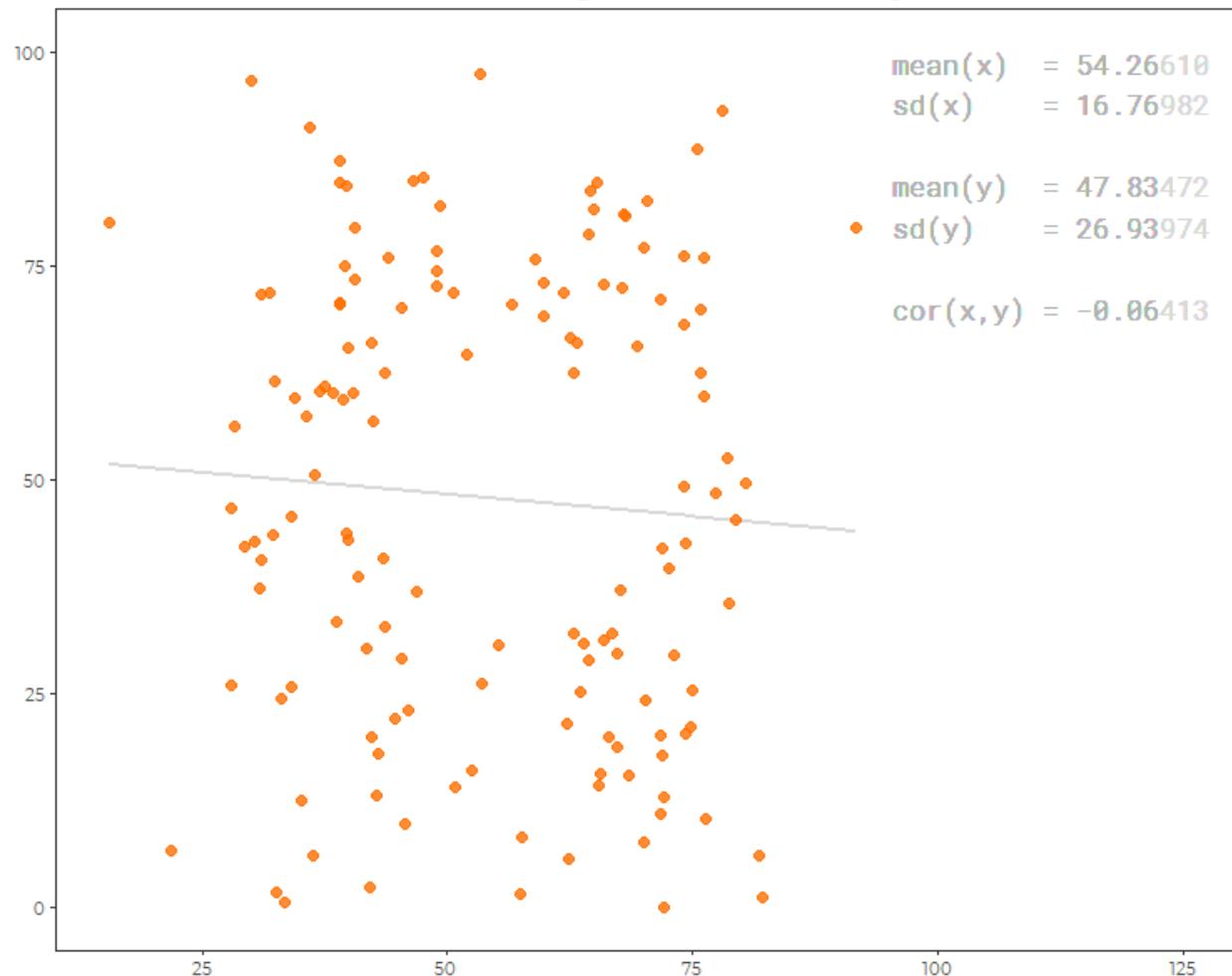
"The Datasaurus Dozen"

based on Anscombe's Quarter and Alberto Cairo's "Datasaurus" (or "Anscombosaurus")

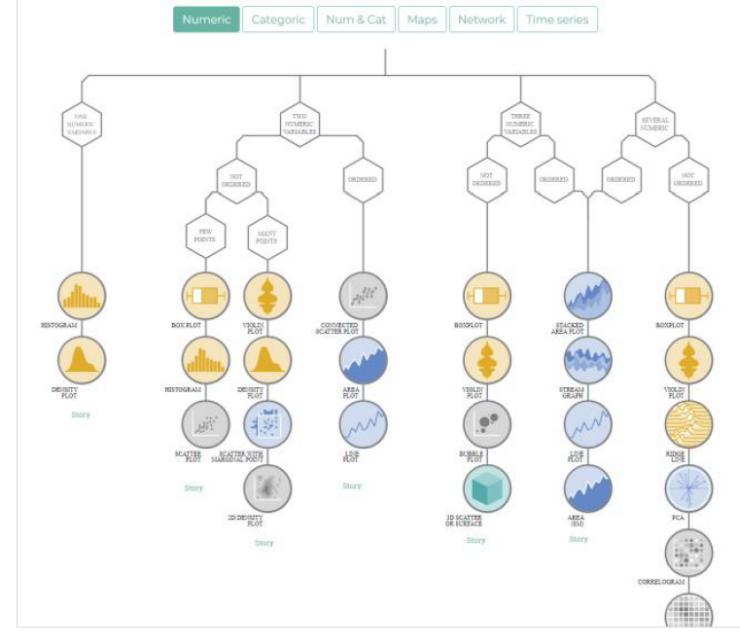


The Datasaurus Dozen

Different datasets – nigh-identical summary statistics



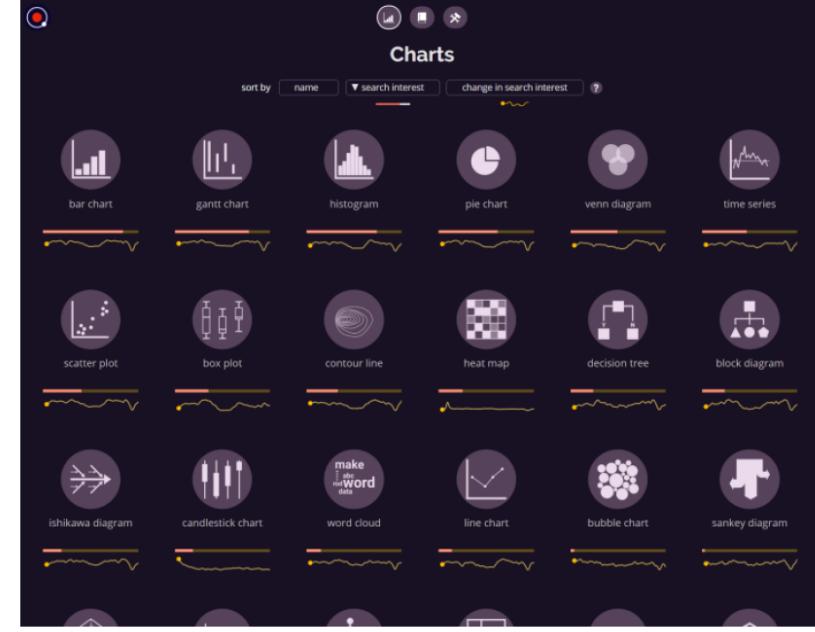
Idea by Alberto Cairo, Justin Matejka & George Fitzmaurice
Visualization by Tom Westlake & Cédric Scherer



data-to-viz.com



datavizproject.com



visualizationuniverse.com/charts



BOXPLOT

Summarize the distribution of numeric variables

[About](#)

A boxplot gives a nice summary of one or several numeric variables. The line that divides the box into 2 parts represents the median of the data. The end of the box shows the upper and lower quartiles. The extreme lines show the highest and lowest value excluding outliers.

[Common Mistakes](#)

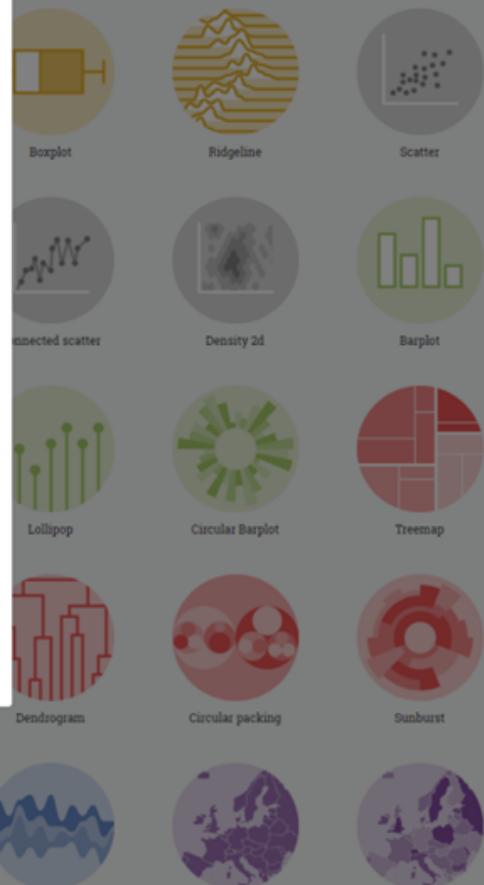
- Boxplot hides the sample size of each group, show it with annotation or box width.
- Boxplot hides the underlying distribution. Use jitter if low number of data points, or use violin with bigger data.
- Order your boxplot by median can make it more insightful.

[Code](#)

[R graph gallery](#) [Python gallery](#) [D3.js gallery](#) [Flourish](#)

[Read More](#)

See the [dedicated page](#).

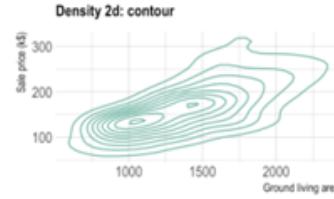
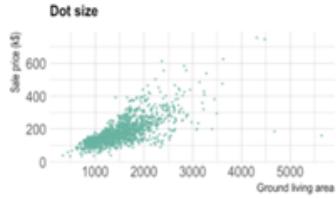


VISIBILITIES
presented in this website.

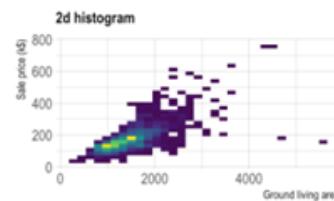
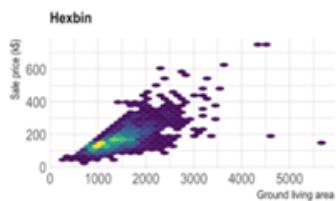
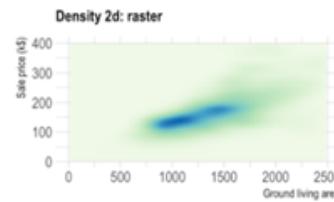
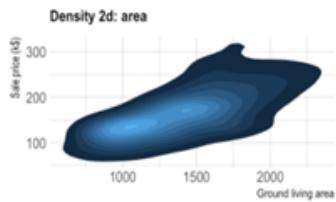
[Part of a whole](#) [Evolution](#) [Map](#) [Flow](#)

Overplotting

The most common pitfall with scatterplot is overplotting: when the sample size gets big, dots are plotted on top of each other what makes the chart unreadable. There are several work around to avoid this issue as describe in this [specific post](#). Here is a summary of the different offered techniques:



CODE



Going further

You can learn more about each type of graphic presented in this story in the dedicated

HIDE

```
# code for all graphics:
p <- data %>%
  ggplot( aes(x=GrLivArea, y=SalePrice/1000)) +
  theme_ipsum() +
  theme(
    plot.title = element_text(size=12)
  ) +
  ylab('Sale price (k$)') +
  xlab('Ground living area')

# Reduce dot size
p1 <- p + geom_point(color="#69b3a2", alpha=0.8, size=0.2) + ggtitle("Dot size")

# Use density estimate
p2 <- p + geom_density2d(color="#69b3a2") + ggtitle("Density 2d: contour")

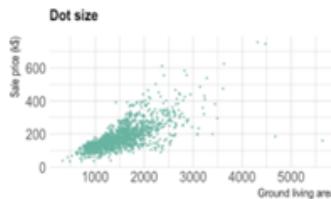
# Use density estimate (area)
p3 <- p + stat_density_2d(aes(fill = ..level..), geom = "polygon") + ggtitle("Density 2d: area") + theme(legend.position="none")

# With raster
p4 <- p +
  stat_density_2d(aes(fill = ..density..), geom = "raster", contour = FALSE) +
  scale_fill_distiller(palette="c", direction=1) +
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0)) +
  theme(
    legend.position="none"
  ) +
  ggtitle("Density 2d: raster") +
  xlim(0,2500) +
  ylim(0,400)

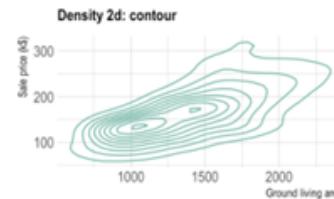
# Hexbin
p5 <- p + geom_hex() +
  scale_fill_viridis() +
  theme(legend.position="none") +
  ggtitle("Hexbin")

# 2d histogram
p6 <- p + geom_bin2d() +
  scale_fill_viridis() +
  theme(legend.position="none") +
  ggtitle("2d histogram")

p1 + p2 + p3 + p4 + p5 + p6 + plot_layout(ncol = 2)
```

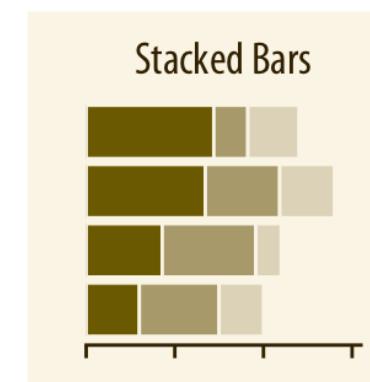
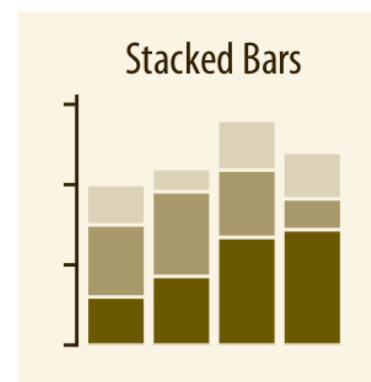
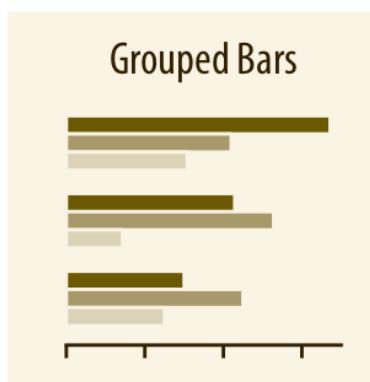
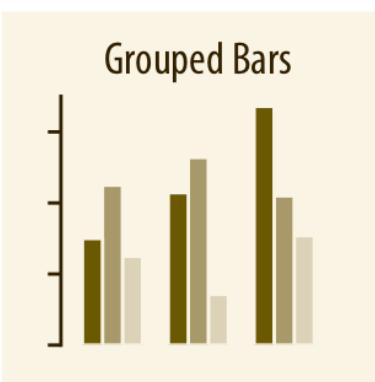
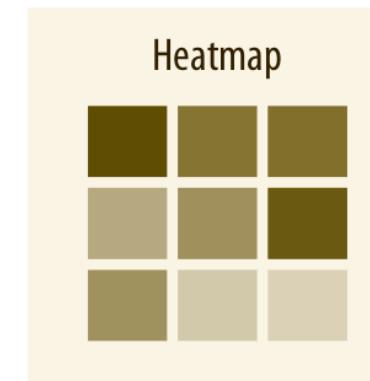
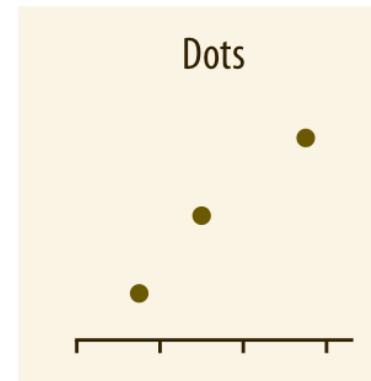
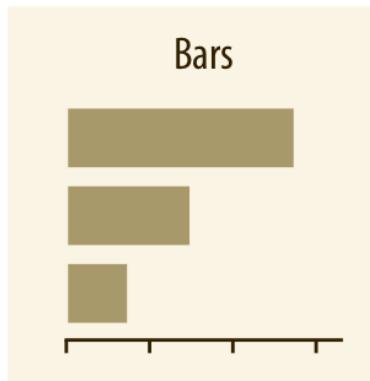
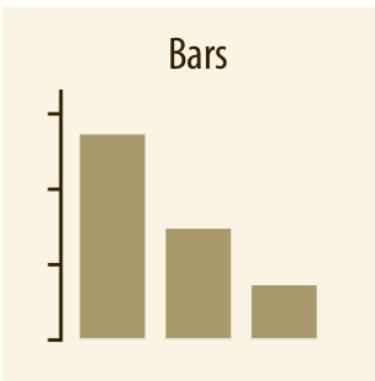


Density 2d: area



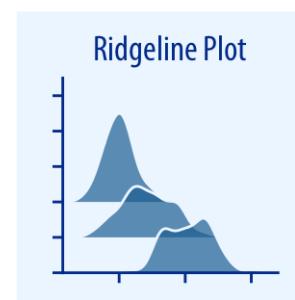
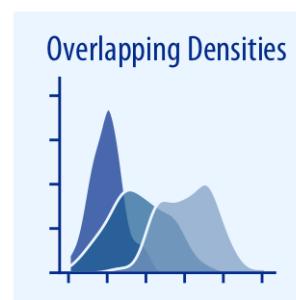
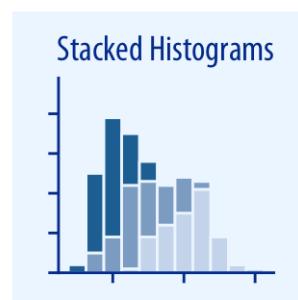
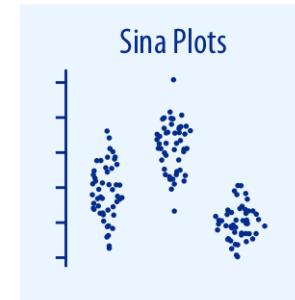
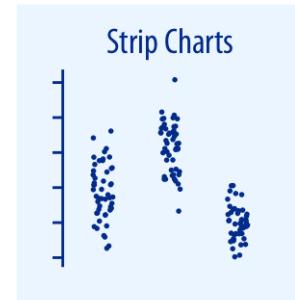
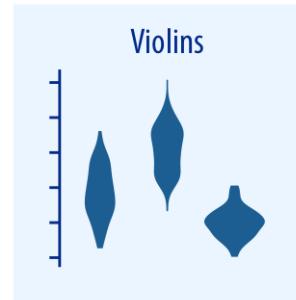
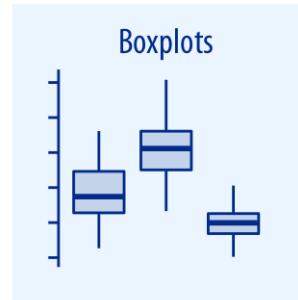
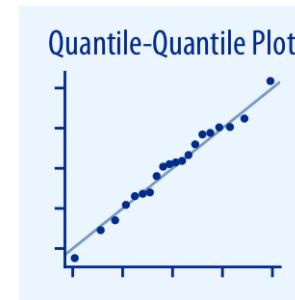
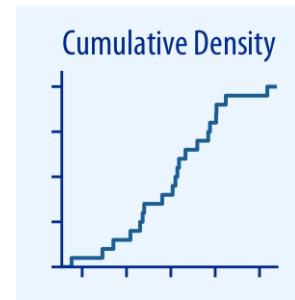
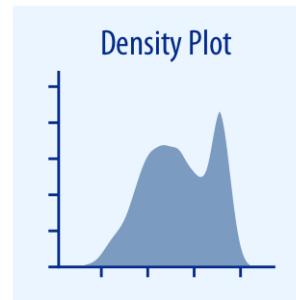
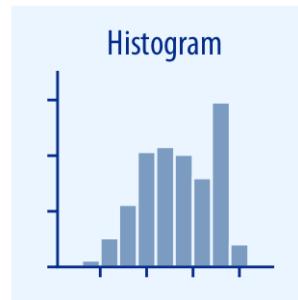
Density 2d: raster

Charts to Visualize Amounts



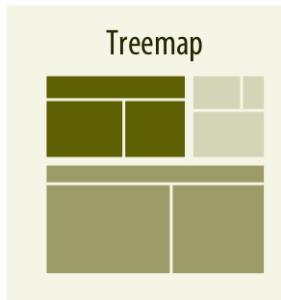
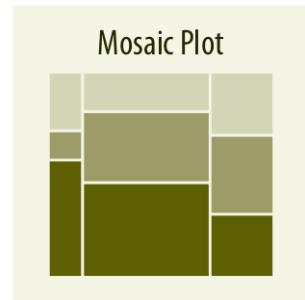
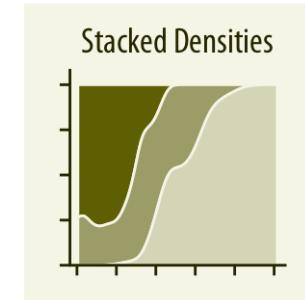
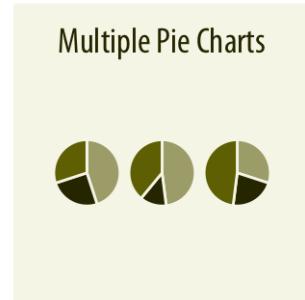
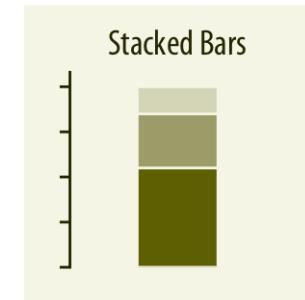
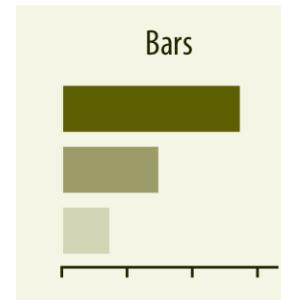
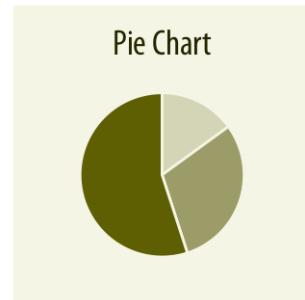
Source: "Fundamentals of Data Visualization" by Claus Wilke

Charts to Visualize Distributions



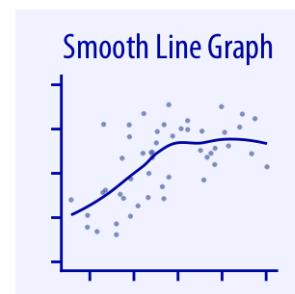
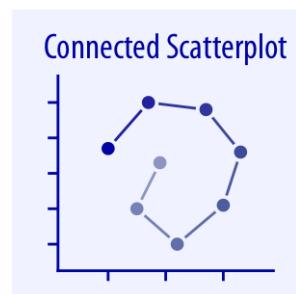
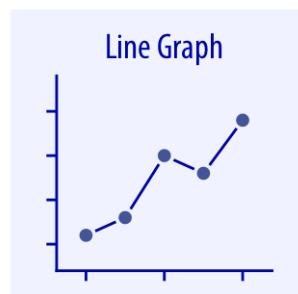
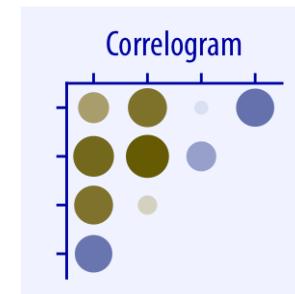
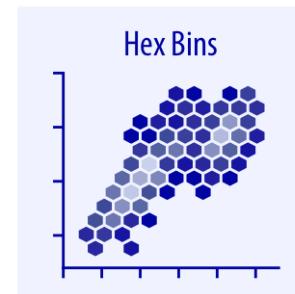
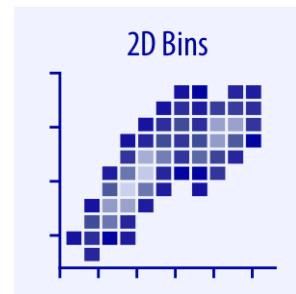
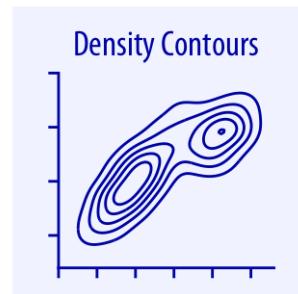
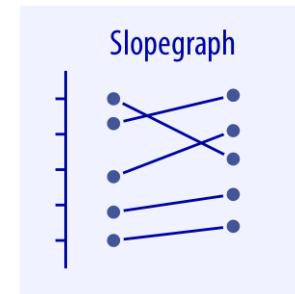
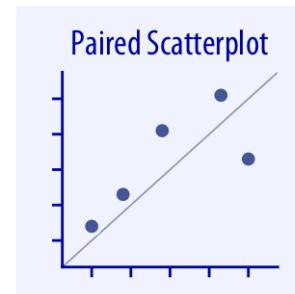
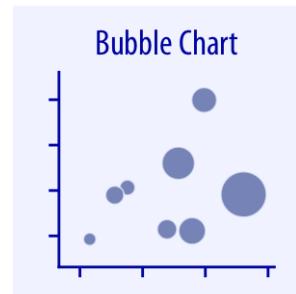
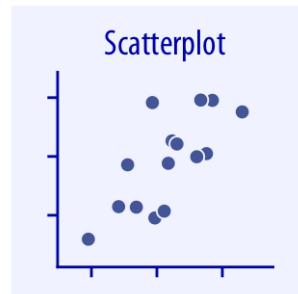
Source: "Fundamentals of Data Visualization" by Claus Wilke

Charts to Visualize Proportions



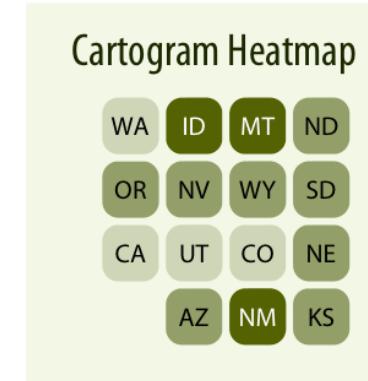
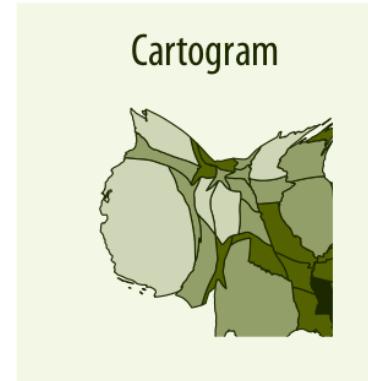
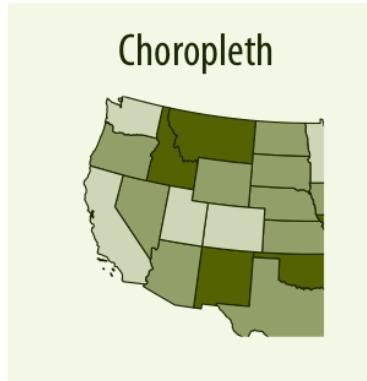
Source: "Fundamentals of Data Visualization" by Claus Wilke

Charts to Visualize x-y Relationships



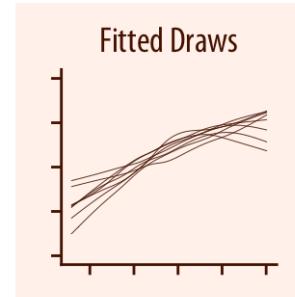
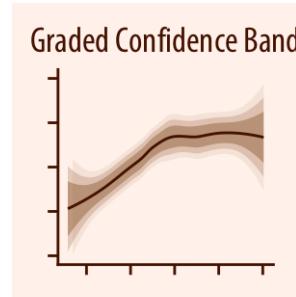
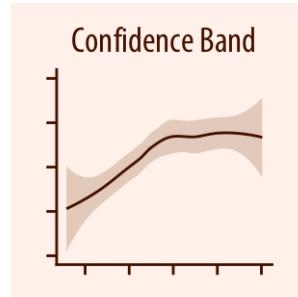
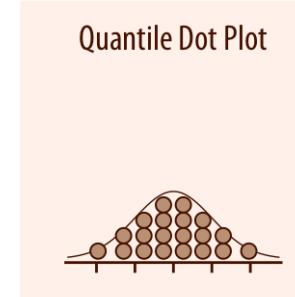
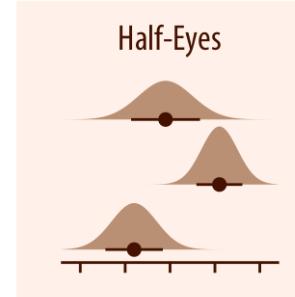
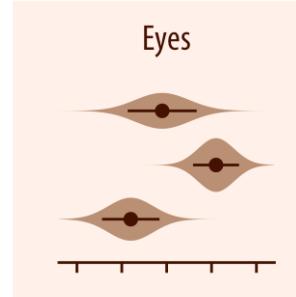
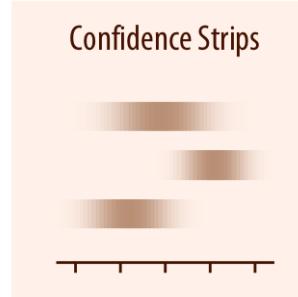
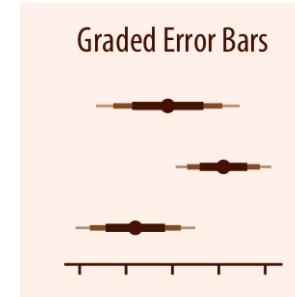
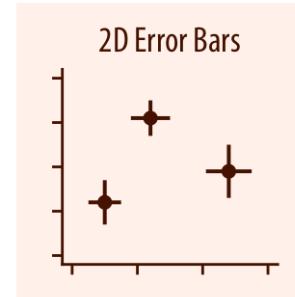
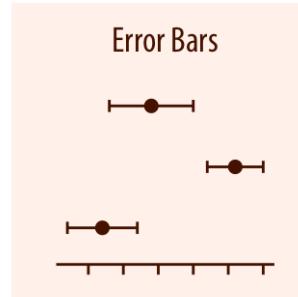
Source: "Fundamentals of Data Visualization" by Claus Wilke

Charts to Visualize Geospatial Data



Source: "Fundamentals of Data Visualization" by Claus Wilke

Charts to Visualize Uncertainty



Source: "Fundamentals of Data Visualization" by Claus Wilke

Your Turn!