# Critical aspects of statistical data analysis:

Experiment design, pre-processing, multivariate methods, visualisation and reporting

Krzysztof Banas
19.10.2022

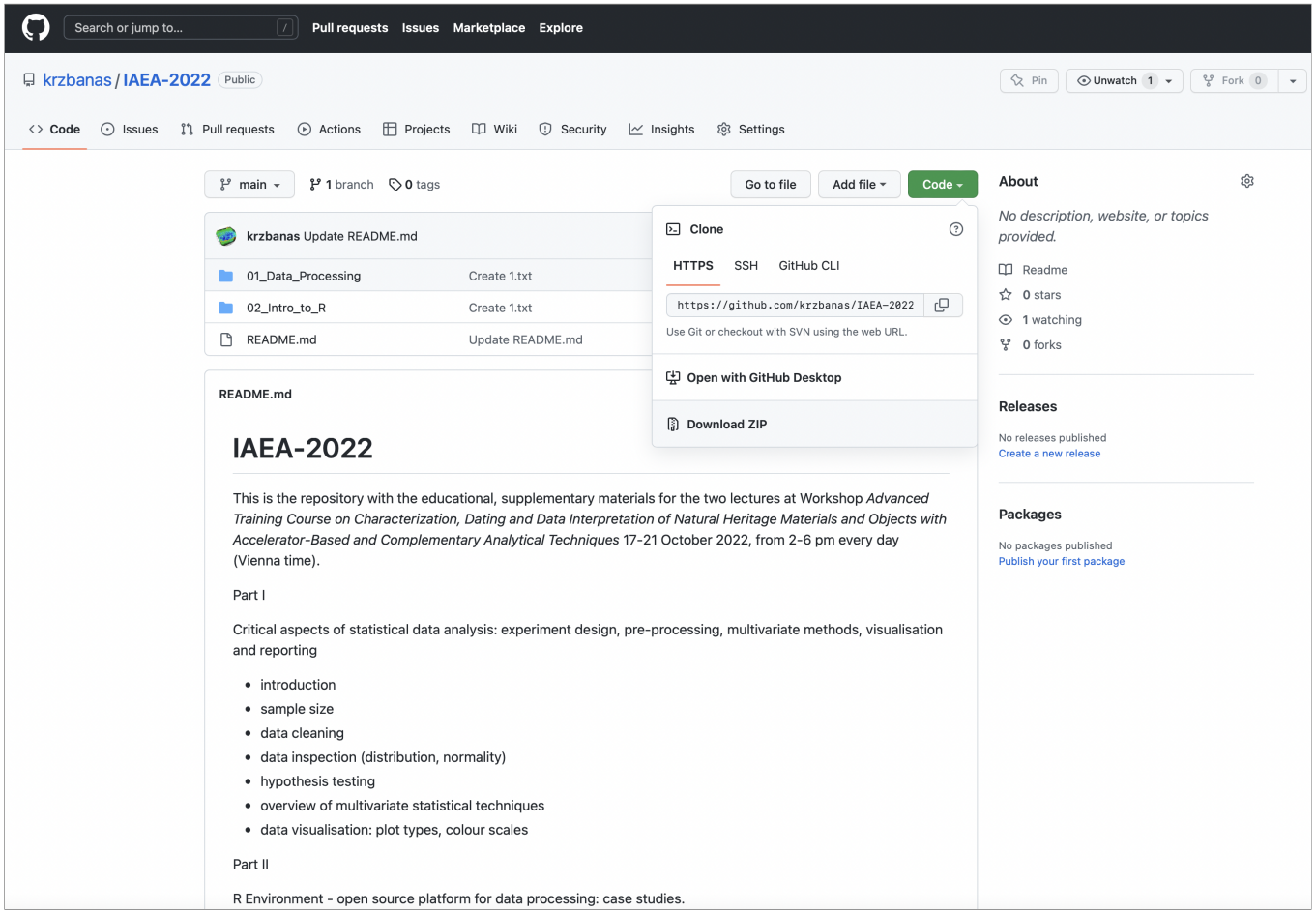
# Overview

- Intro 👋
- Workshop materials ⬇️
- Break 🕘
- Lecture 3.1 plan 📋

## Lecture 3.1 plan

1. Scientific Method
2. Data science
3. Types of visualisation
4. Normal data
5. Can we trust mean value?
6. Hypothesis testing
7. Overview of multivariate statistical techniques

# Accessing workshop materials




**https://github.com/krzbanas/IAEA-2022**
click big green Code button and select “Download ZIP”

## Scientific Method

1. Define a question
2. Gather information and resources (observe)
3. **Form an explanatory hypothesis**
4. **Test the hypothesis by performing an experiment and collecting data in a reproducible manner**
5. **Analyze the data**
6. **Interpret the data and draw conclusions that serve as a starting point for a new hypothesis**
7. Publish results
8. Retest (frequently done by other scientists)

## Critical Aspects

1. Data Wrangling & Cleaning
2. Data Preprocessing & Feature Engineering
3. Data Visualization
4. Machine Learning
5. Time Series and Forecasting
6. Text Analysis
7. Functional Programming
8. Reporting
9. Applications and deployment

## 1. Data Wrangling & Cleaning

- Working with outliers
- Missing data
- Reshaping data
- Aggregation
- Filtering
- Selecting
- Calculating

## 2. Data Preprocessing & Feature Engineering

- Preparing data for machine learning
- Engineering features (dates, text, aggregates)

### 3. Data Visualization

- Interactive and static visualizations
- For data exploration and for presentation

## 4. Machine Learning

- Supervised classification
- Supervised regression
- Unsupervised clustering
- Dimensionality reduction
- XGBoost, SVM, Random Forest, GLM
- K-Means
- UMAP

## 5. Time Series and forecasting

- Working with date/datetime data
- aggregating
- transforming
- visualising time series
- ARIMA

## 6. Text

- Working with text data
- Strings
- Machine learning
- Text features

## 7. Functional Programming

- Making reusable functions
- Sourcing code
- Iteration
- Loops and mapping

## 8. Reporting

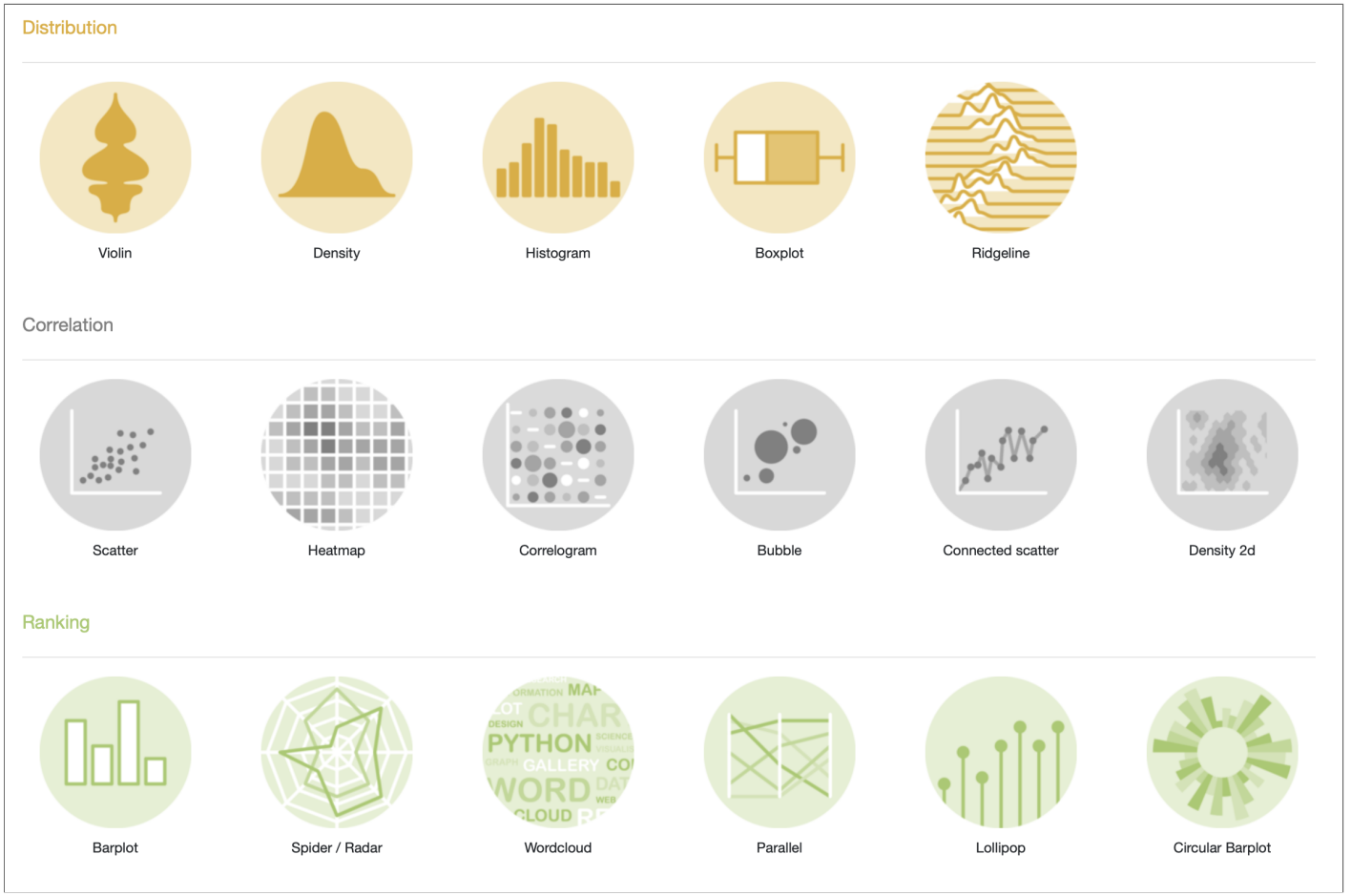
- RMarkdown
- Iteractive HTML
- Static PDF

## 9. Applications and deployment

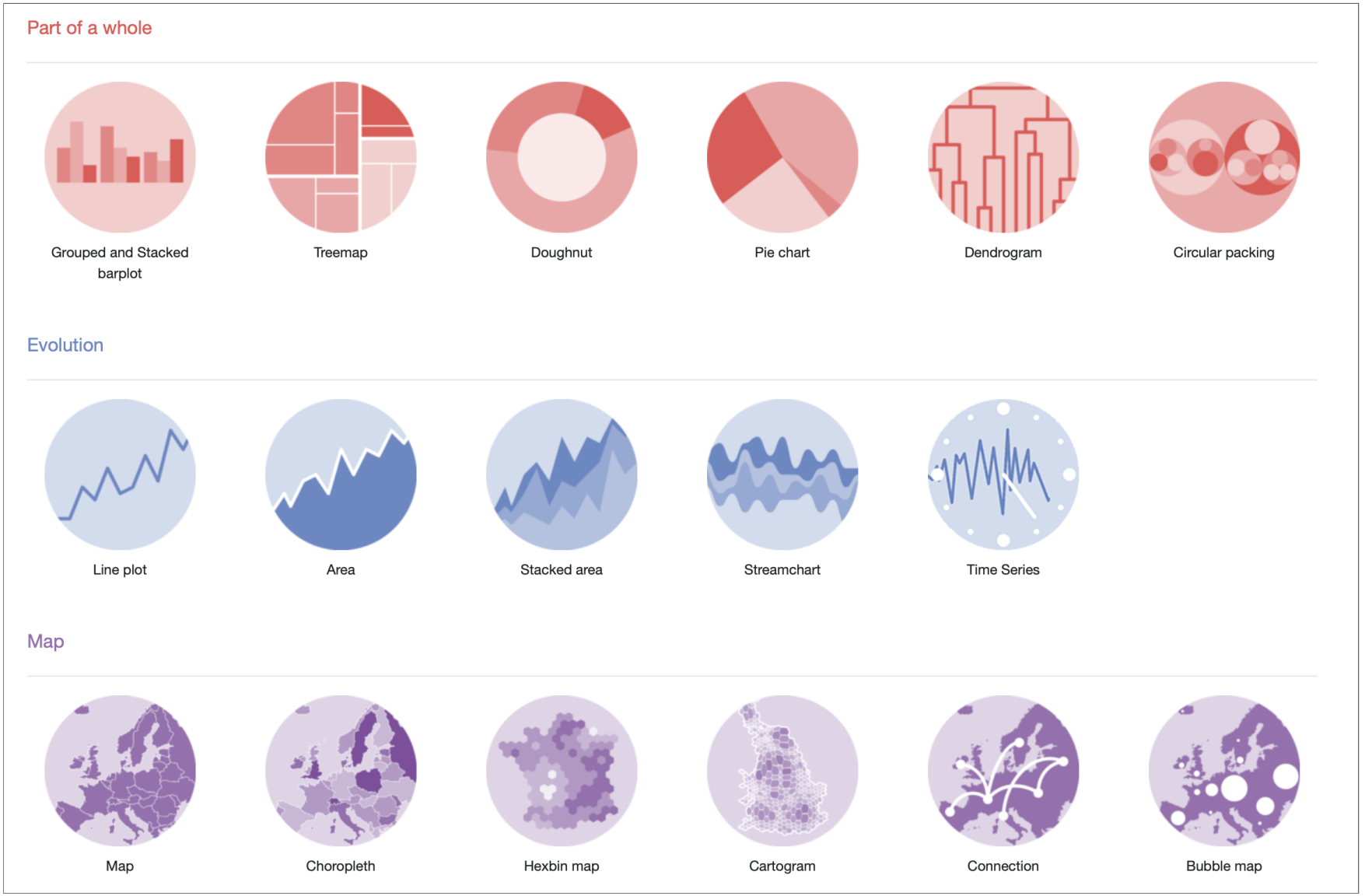
- Building Shiny web applications
- Flexdashboard
- Bootstrap
- Deployment
- Cloud (AWS, Azure, GCP),
- Docker
- Git

# Data visualisation

# Plot types Part 1

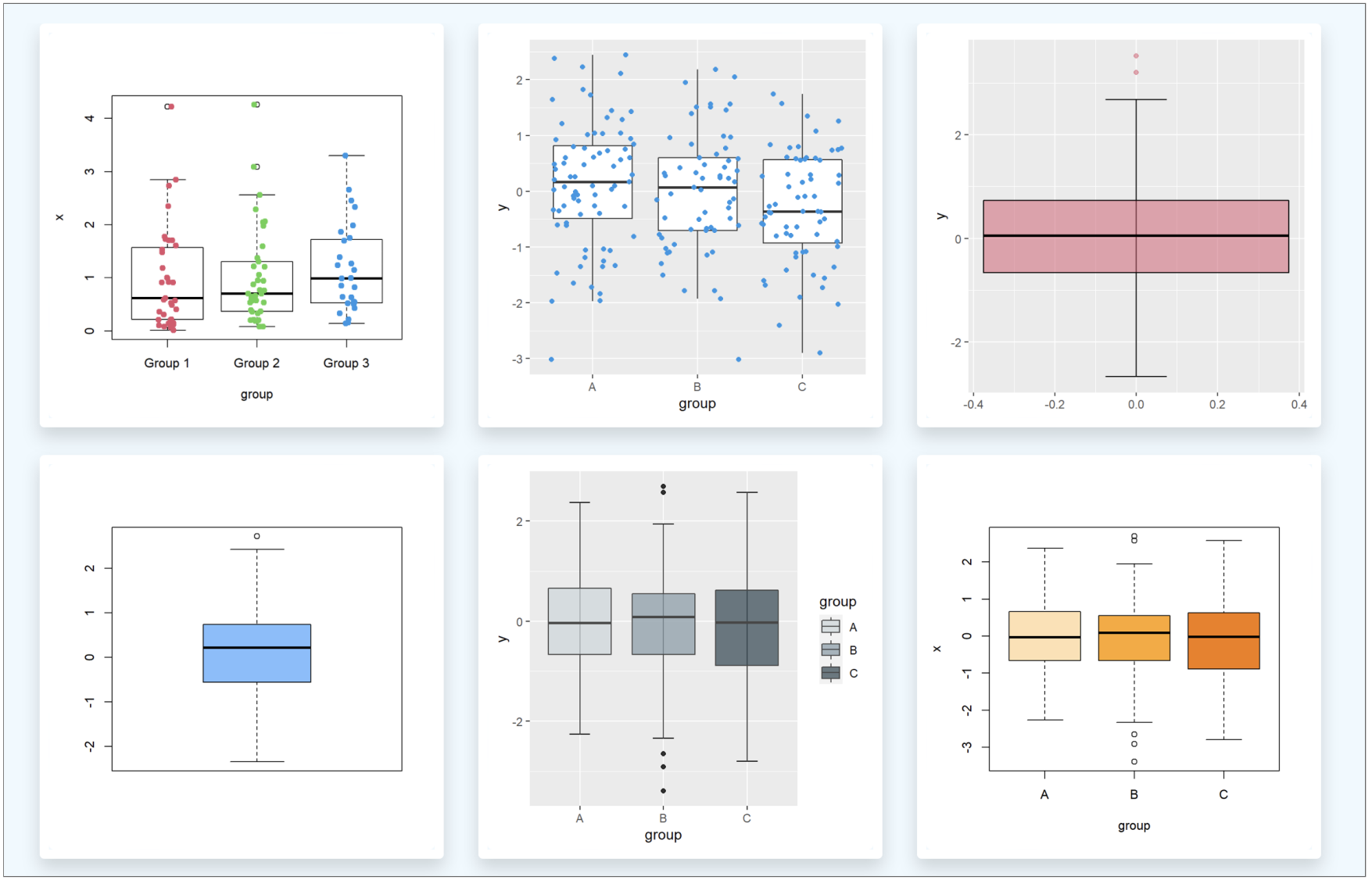


# Plot types Part 2



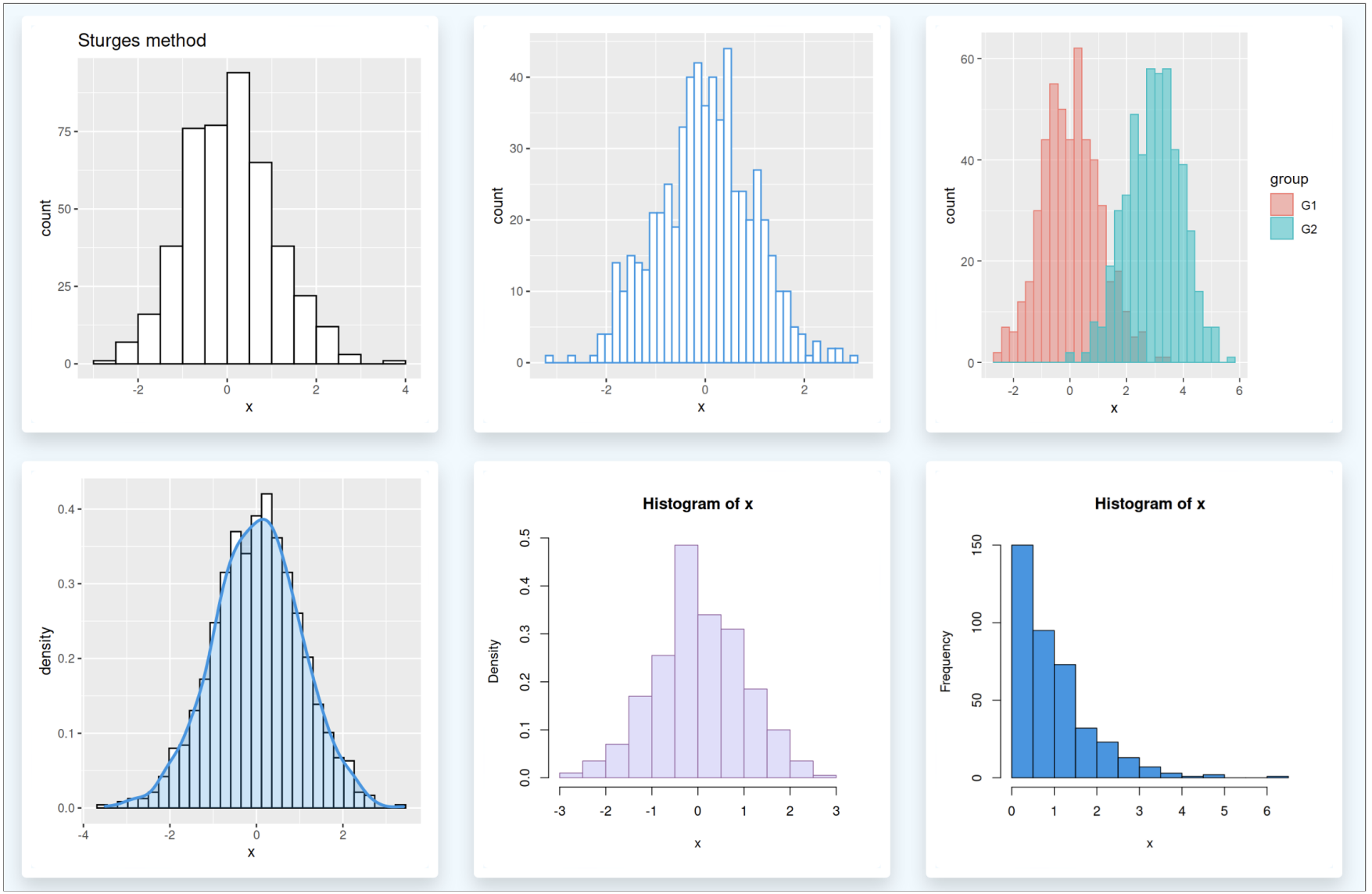

{fig-align="center"}

# Distributions 1



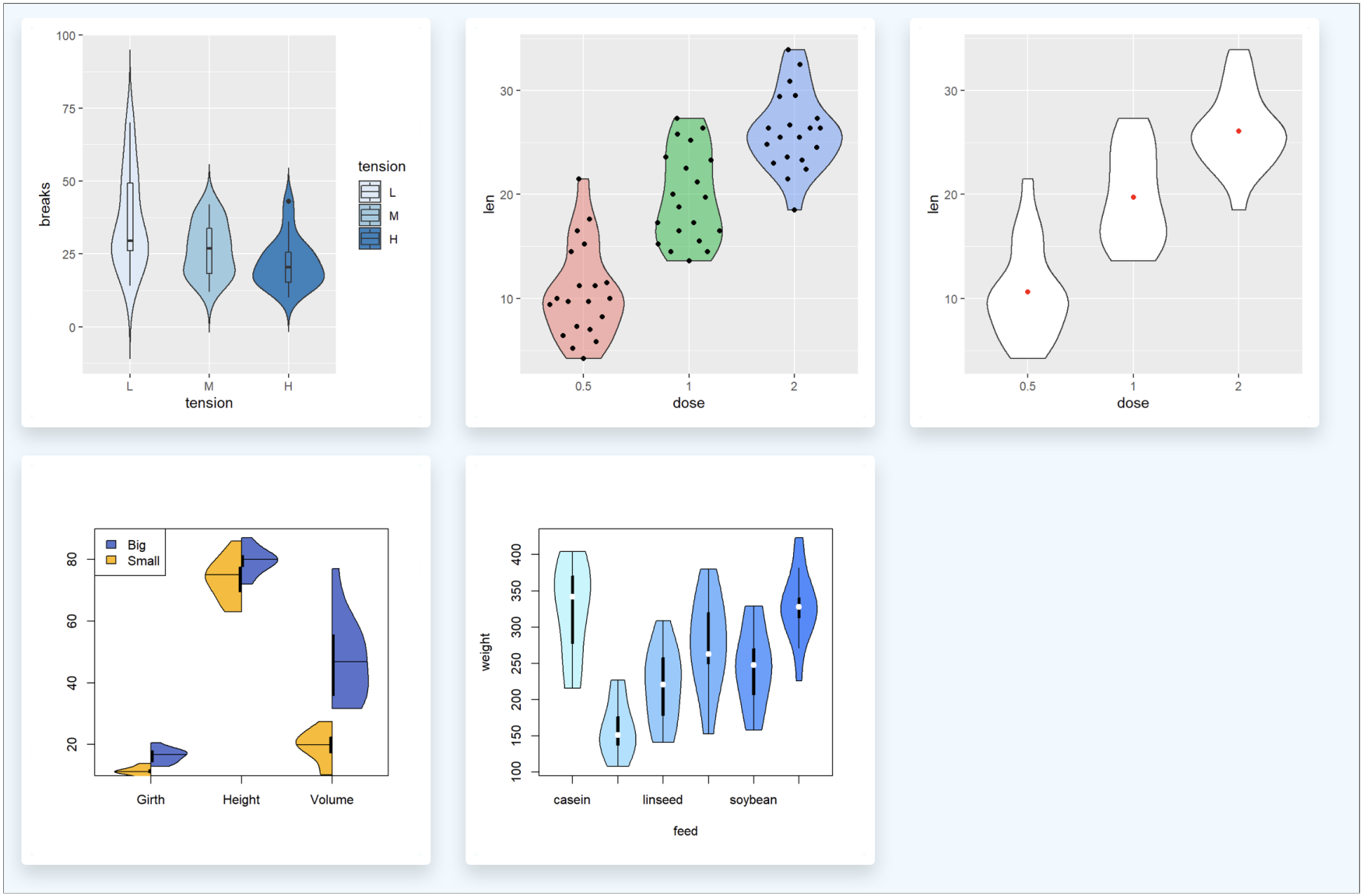

{fig-align="center"}

# Distributions 2



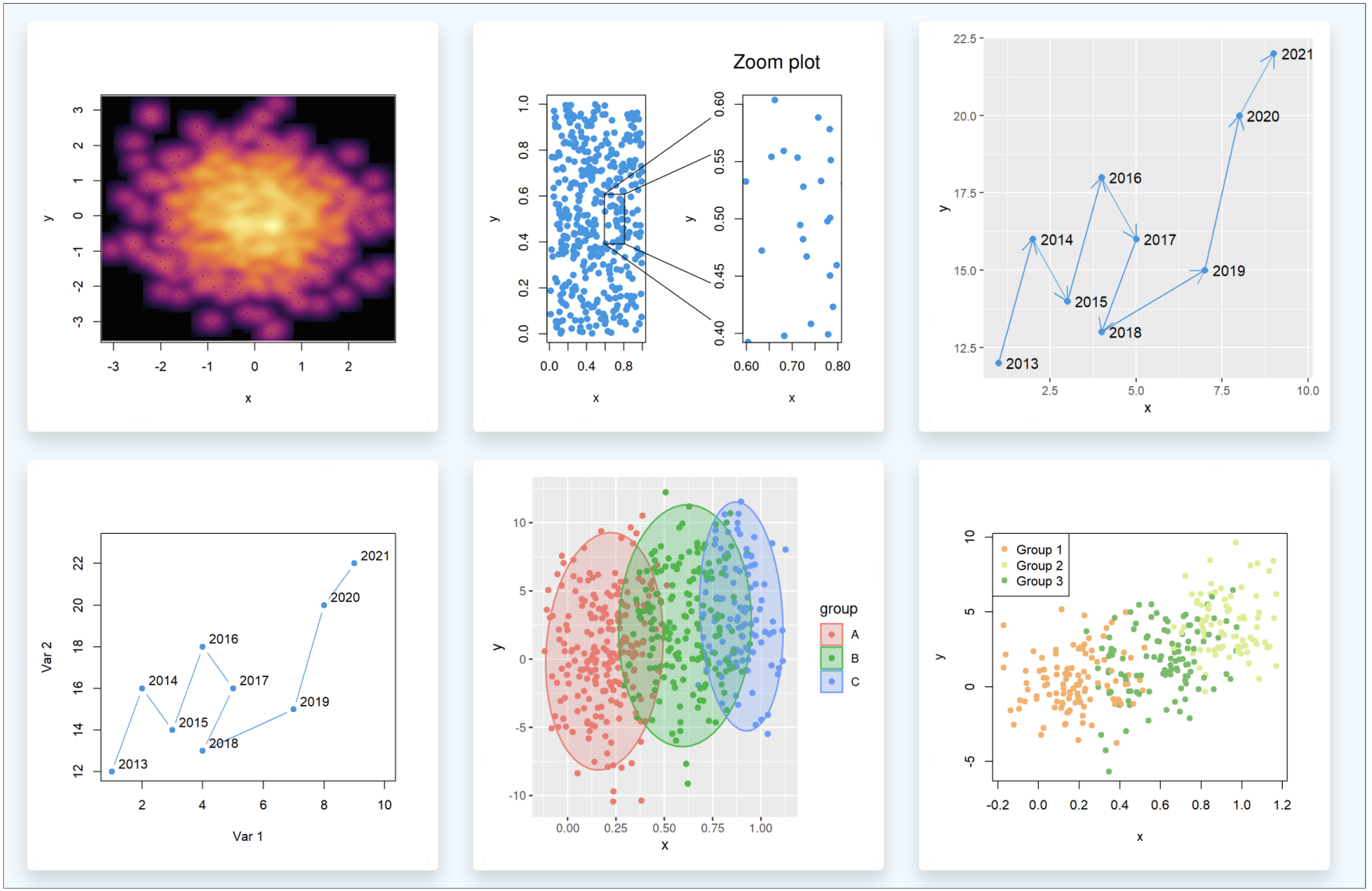

{fig-align="center"}

# Distributions 3



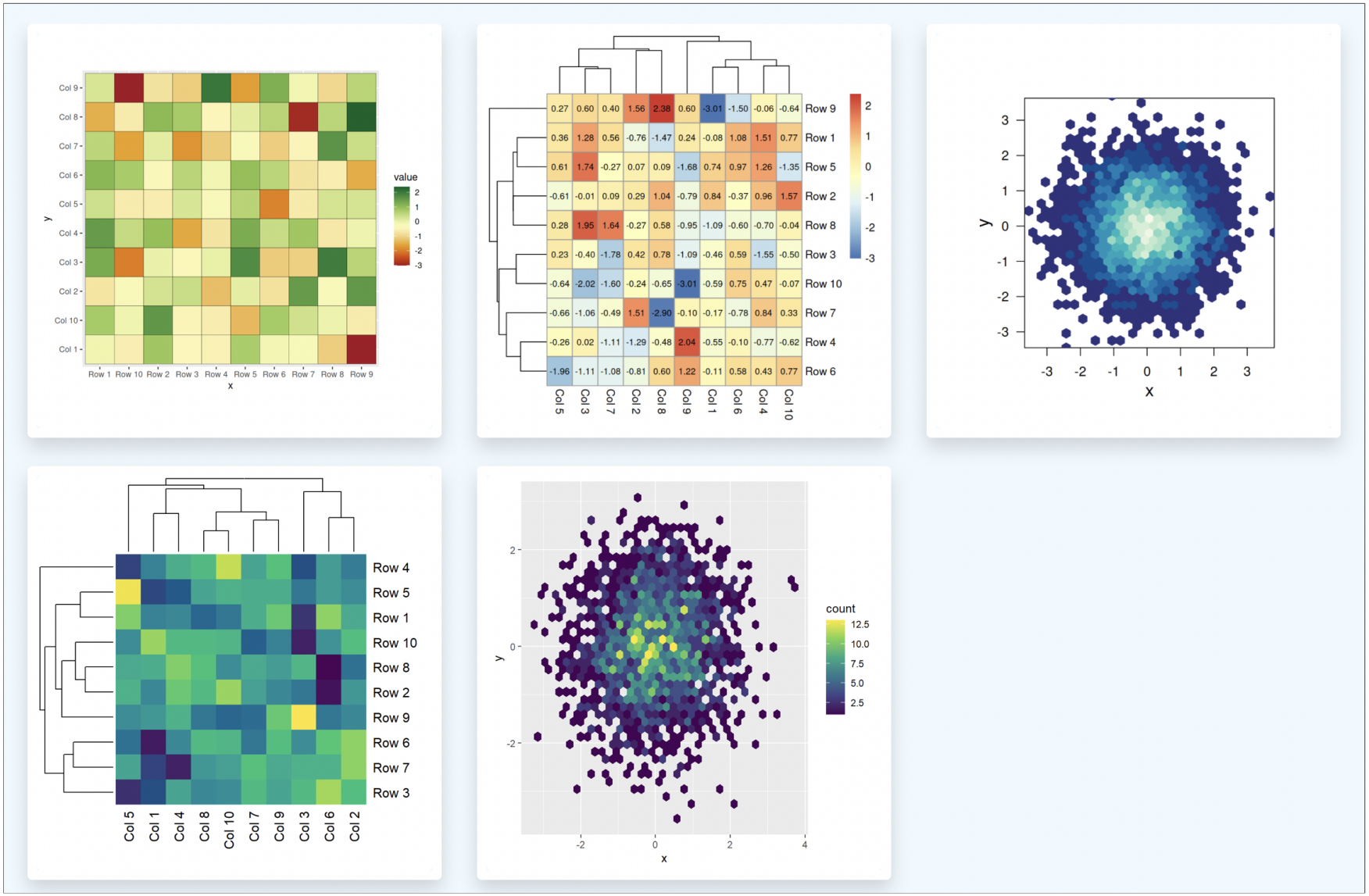

{fig-align="center"}

# Correlation 1



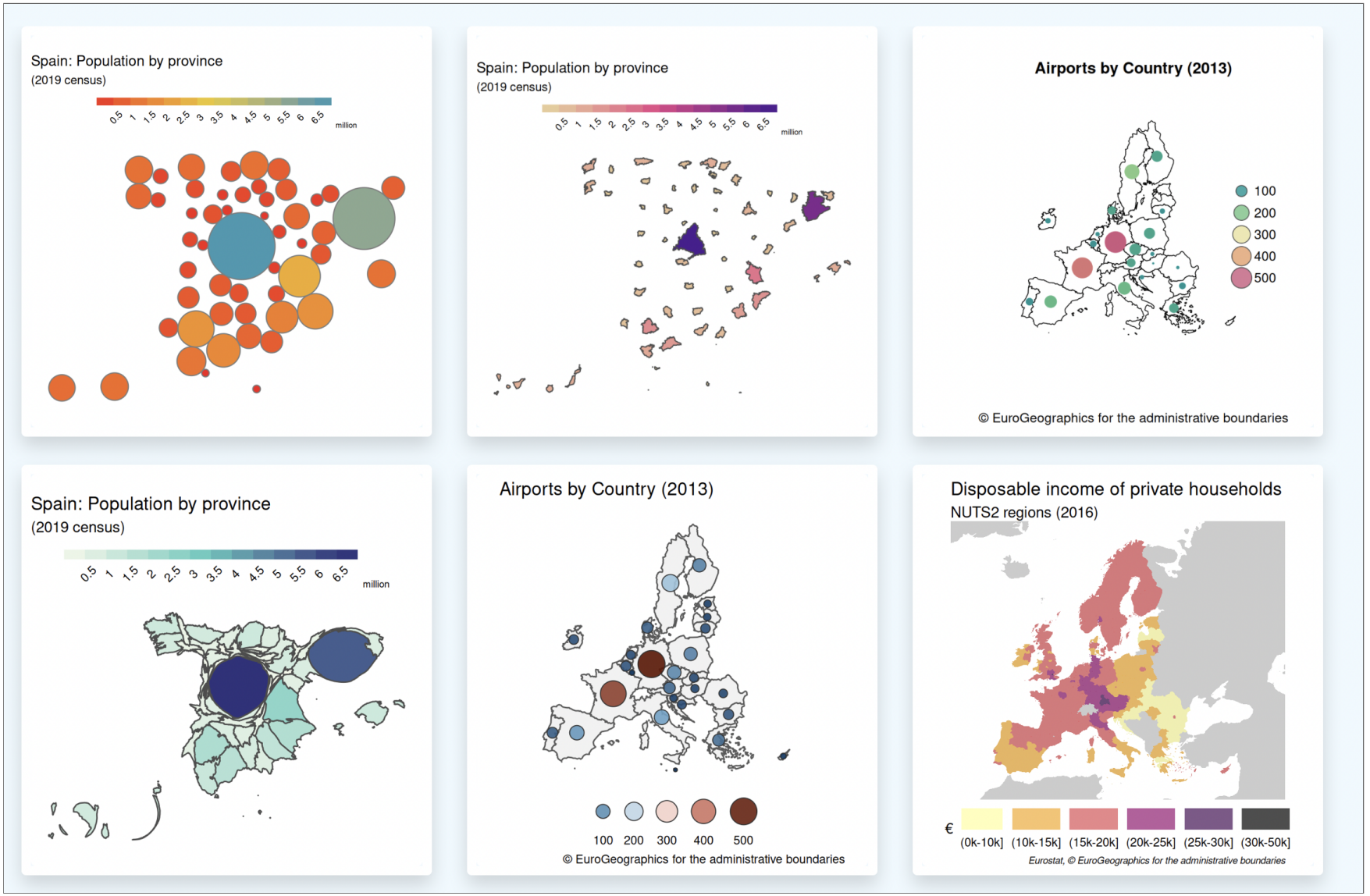

{fig-align="center"}

# Correlation 2



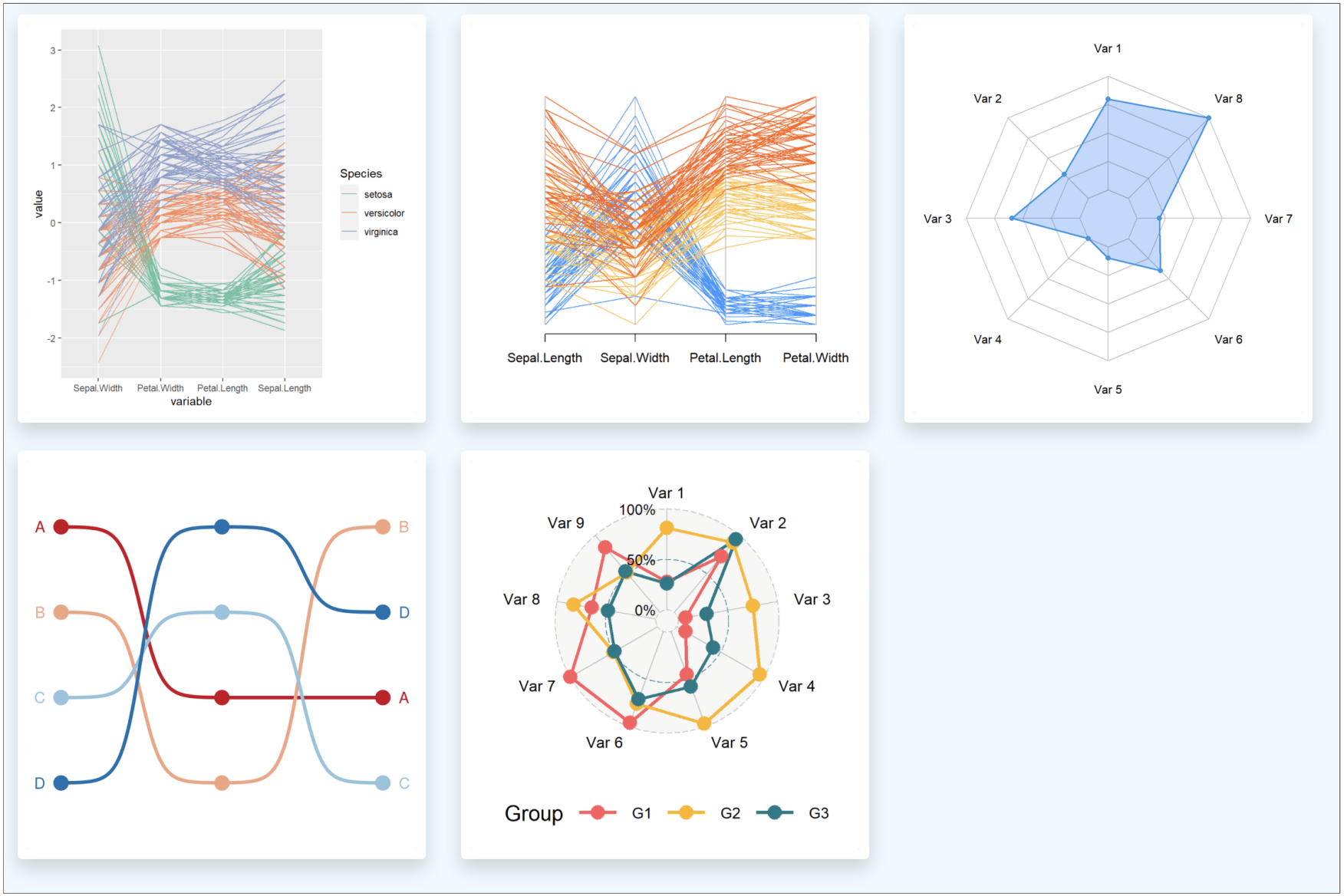

{fig-align="center"}

# Spatial



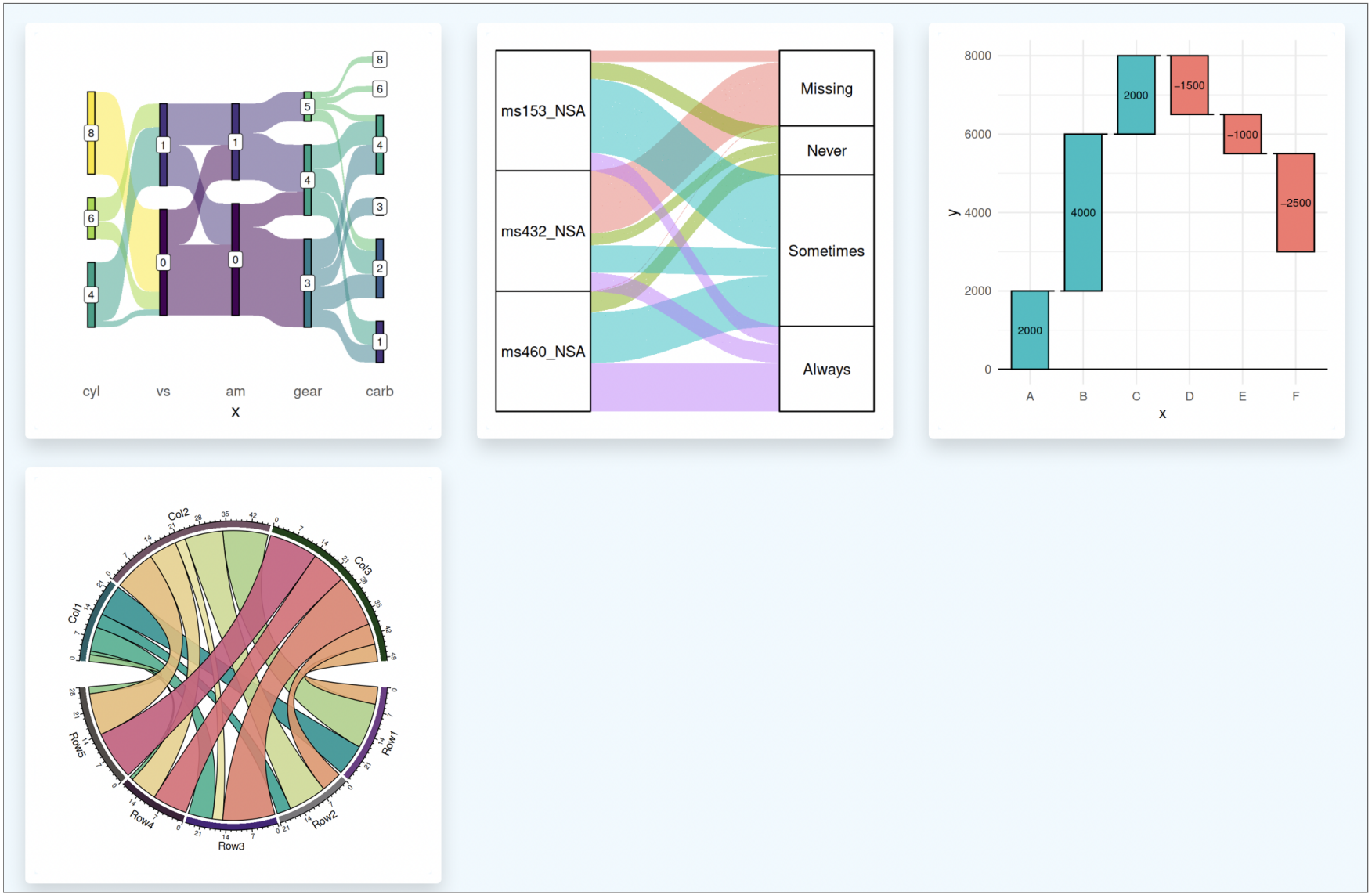

{fig-align="center"}

# Ranking



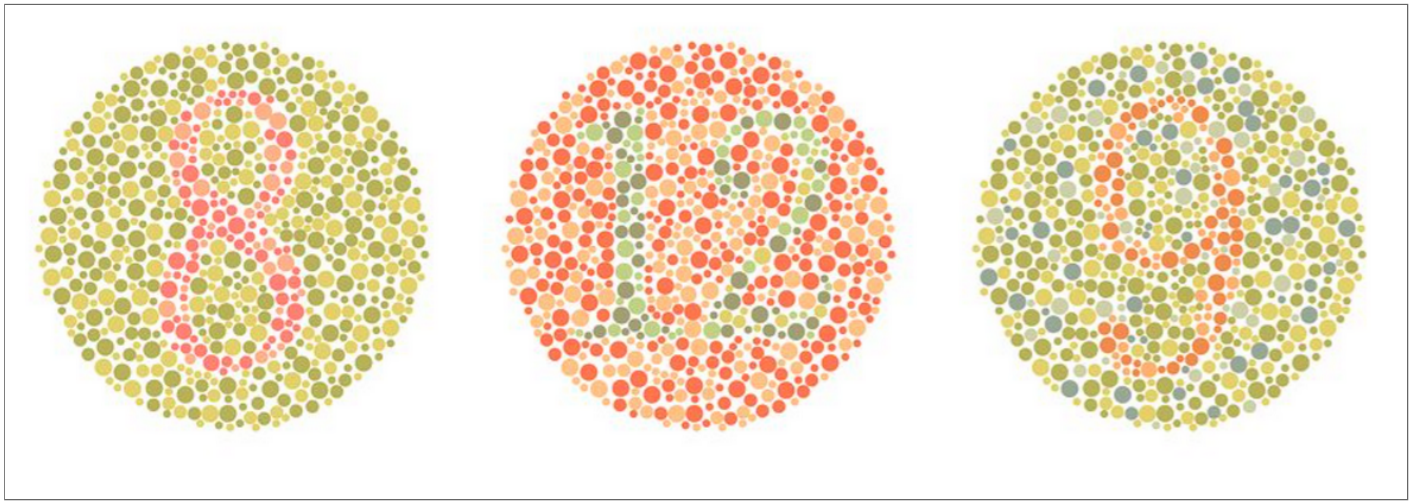

{fig-align="center"}

## Flow

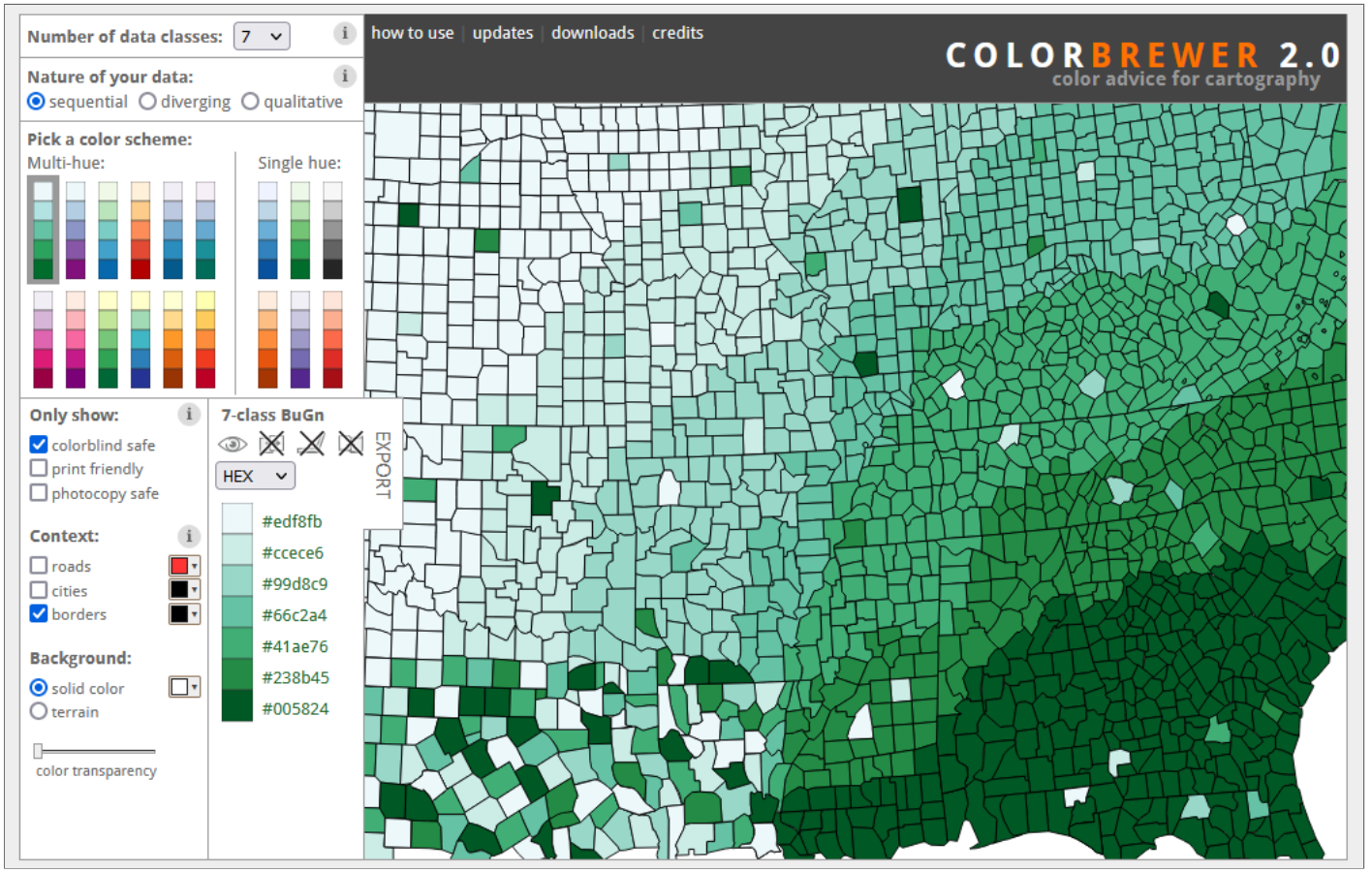



{fig-align="center"}

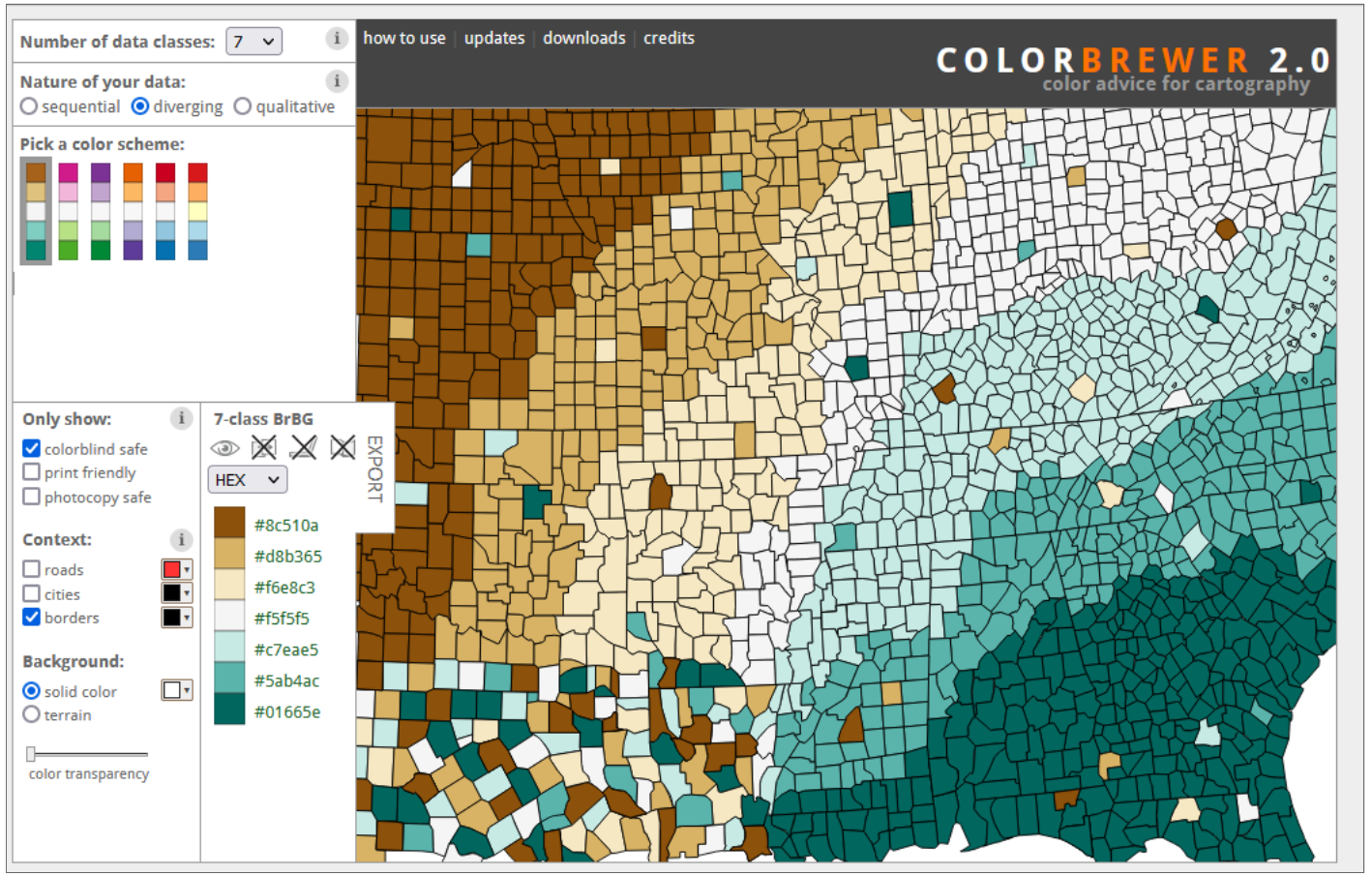
# Data visualisation: colour scales



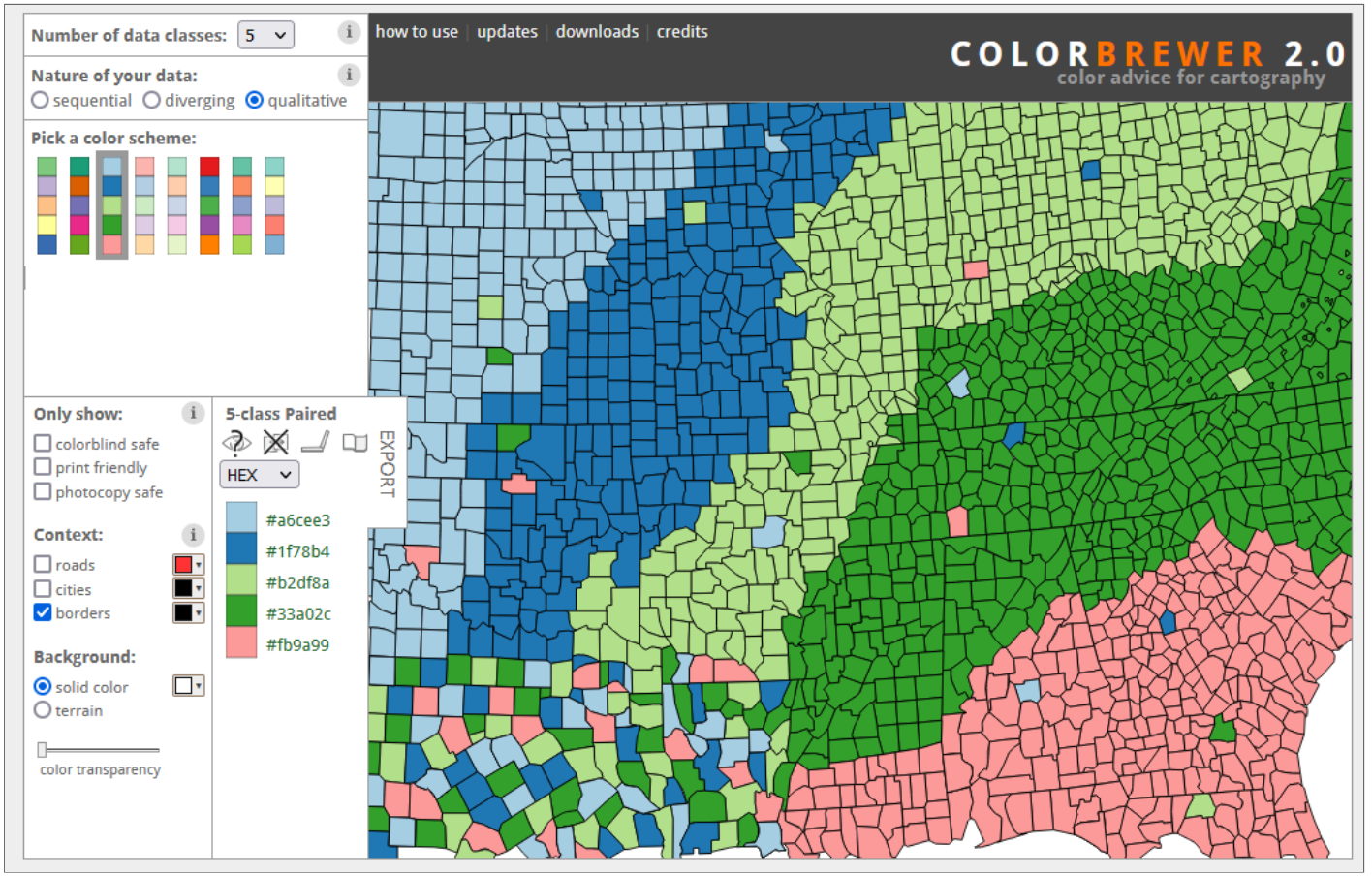
# Sequential


Number of data classes: 7
Nature of your data:
sequential
diverging
qualitative
Pick a color scheme:
Multi-hue:
Single hue:
Only show:
colorblind safe
print friendly
photocopy safe
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency
7-class BuGn
HEX
EXPORT
#edf8fb
#ccece6
#99d8c9
#66c2a4
#41ae76
#238b45
#005824
how to use
updates
downloads
credits
COLORBREWER 2.0
color advice for cartography

# Diverging


Number of data classes: 7
Nature of your data:
sequential
diverging
qualitative
Pick a color scheme:
Only show:
colorblind safe
print friendly
photocopy safe
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency
7-class BrBG
HEX
EXPORT
#8c510a
#d8b365
#f6e8c3
#f5f5f5
#c7eae5
#5ab4ac
#01665e
how to use
updates
downloads
credits
COLORBREWER 2.0
color advice for cartography

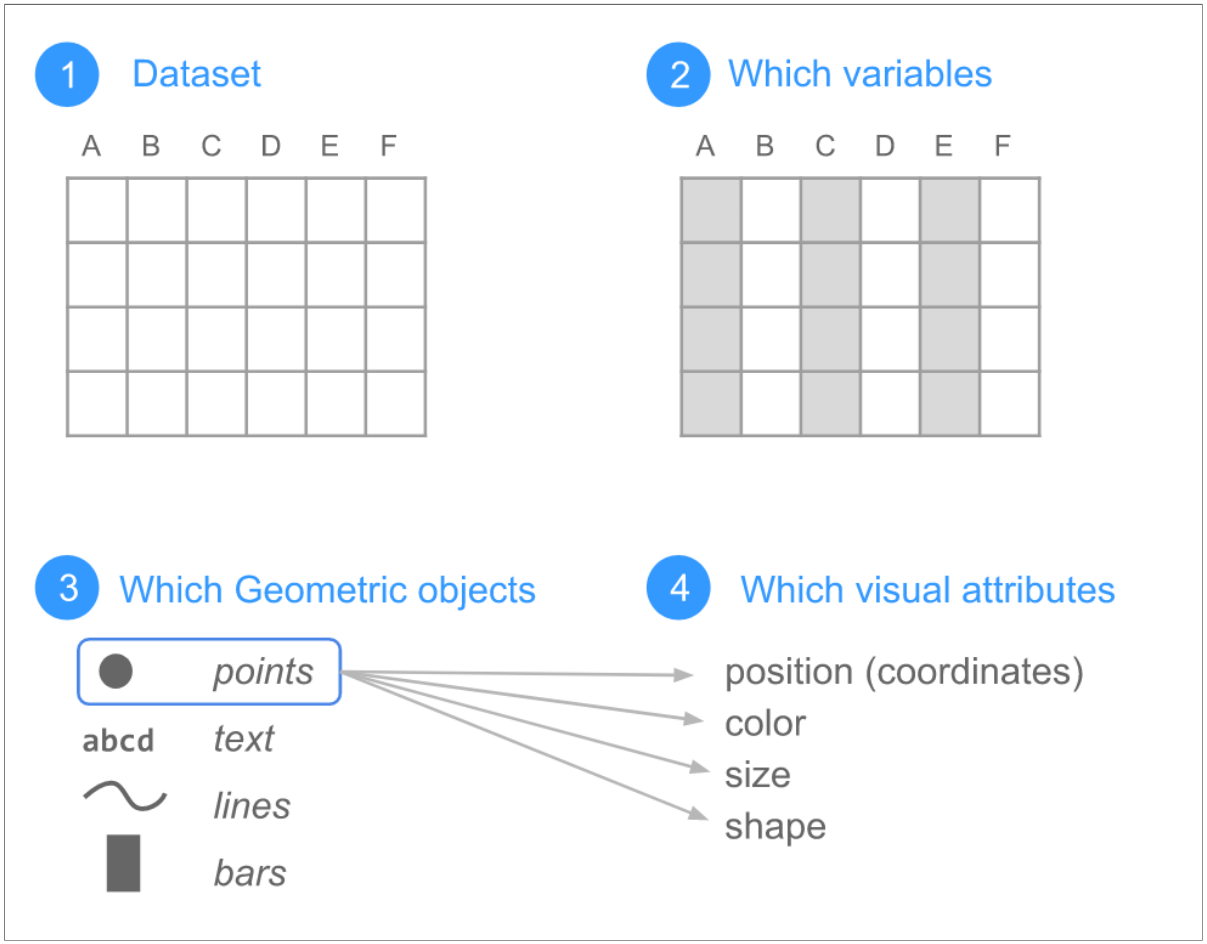
# Qualitative



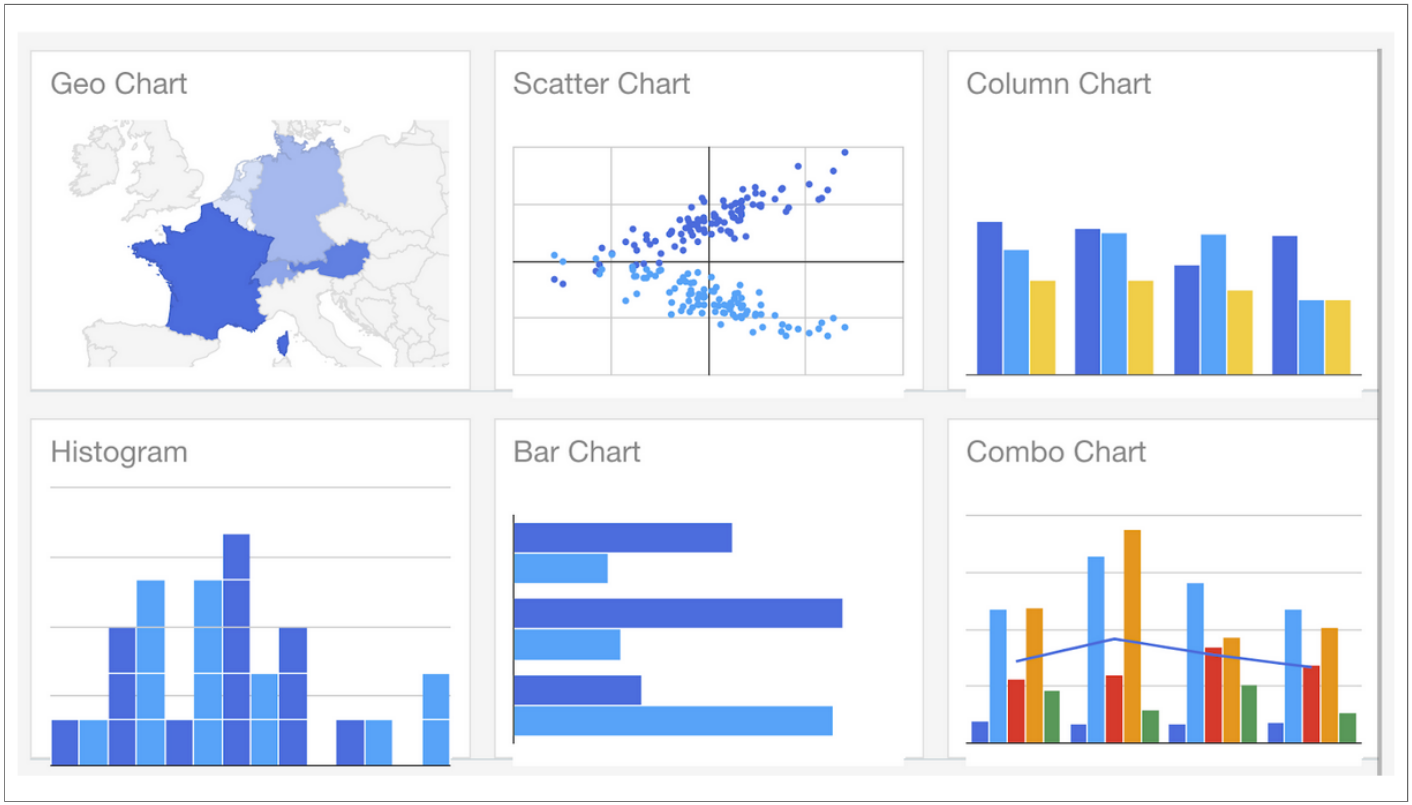
## Data Visualization

- core idea : mapping data to geometric objects and their visual attributes (points, lines, bars, polygons ) (position, shape, colour hue, colour luminance/saturation)
- common analytical tasks
- visual perception
- effective charts suggestions

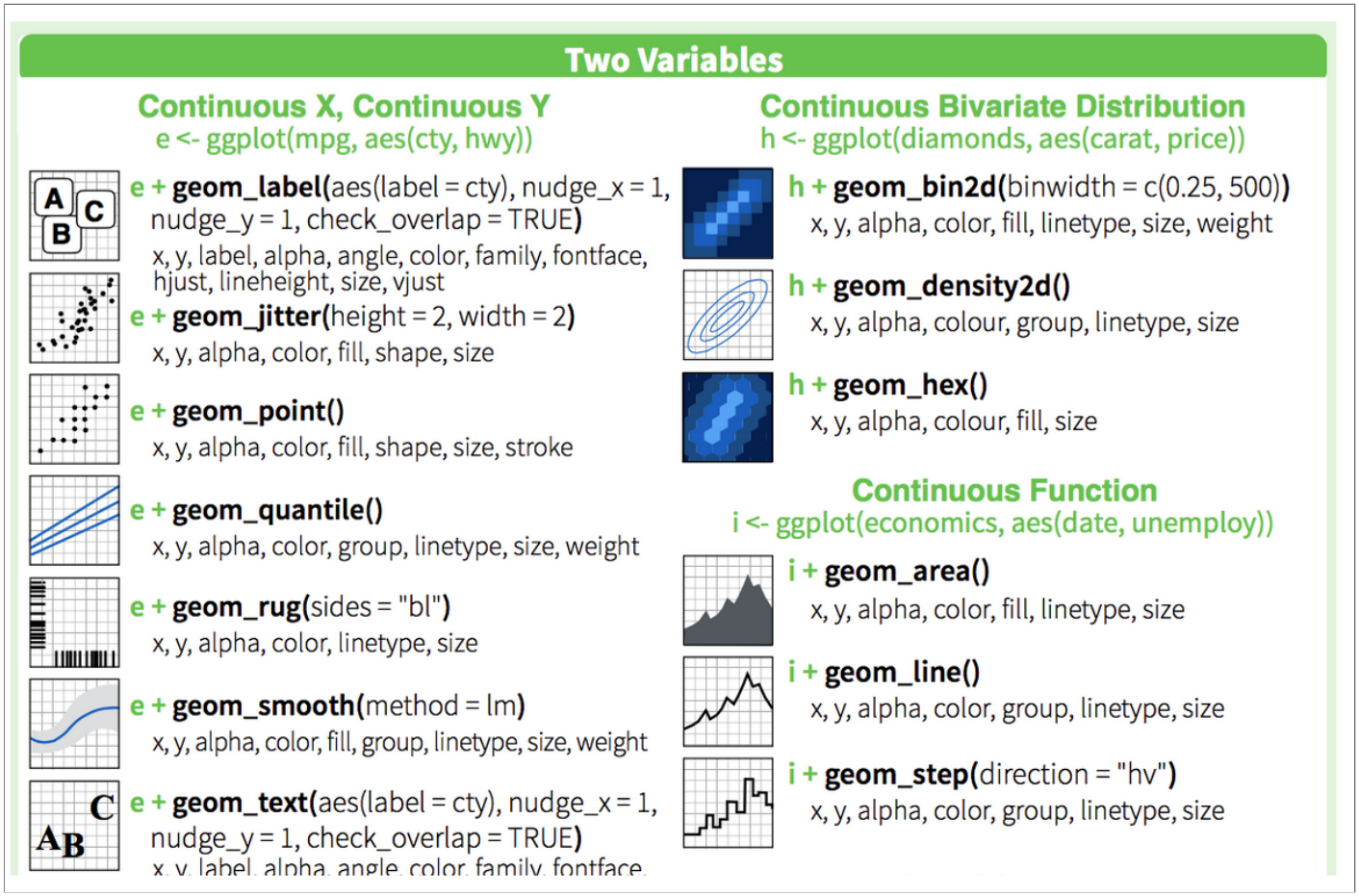
## Visualization Concept



# Some examples - Google Charts



## Some examples - Google Charts



## Data Analysis Cycle

- Data Preparation
- Core Analysis
- Reporting

## Graphics for exploration

- for data verification and understanding
- data analyst is the only consumer
- typically quick and without taking care of visual appearance
- short lifespan

## Graphics for communication

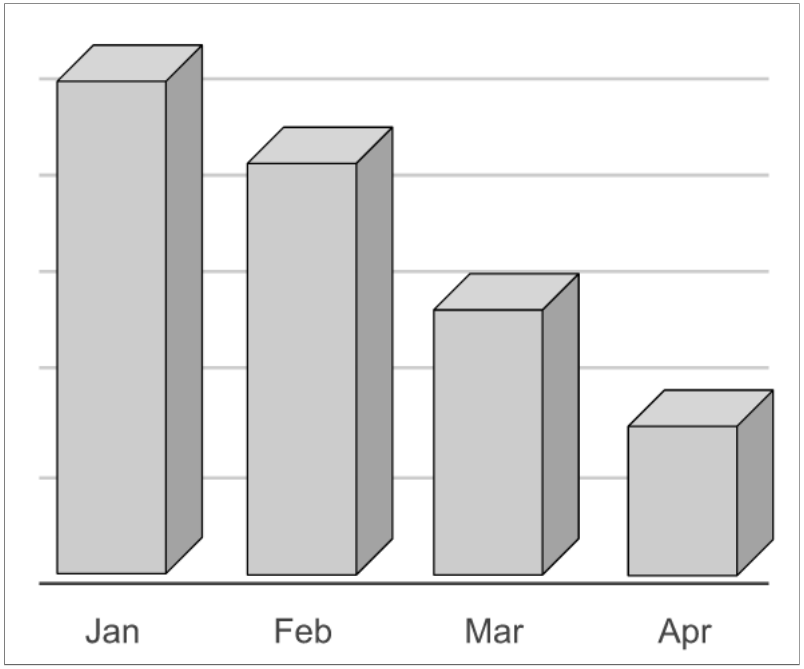
- for presenting data
- understandable by others
- visual appearance and design principles
- require a lot of iterations before final version

## What to consider?

- How many variables (1, 2, 3 or more) ?
- What type of variables (quantitative, qualitative, time/date, location)?
- What do you want to see (variation, increasing pattern, outliers, noise)

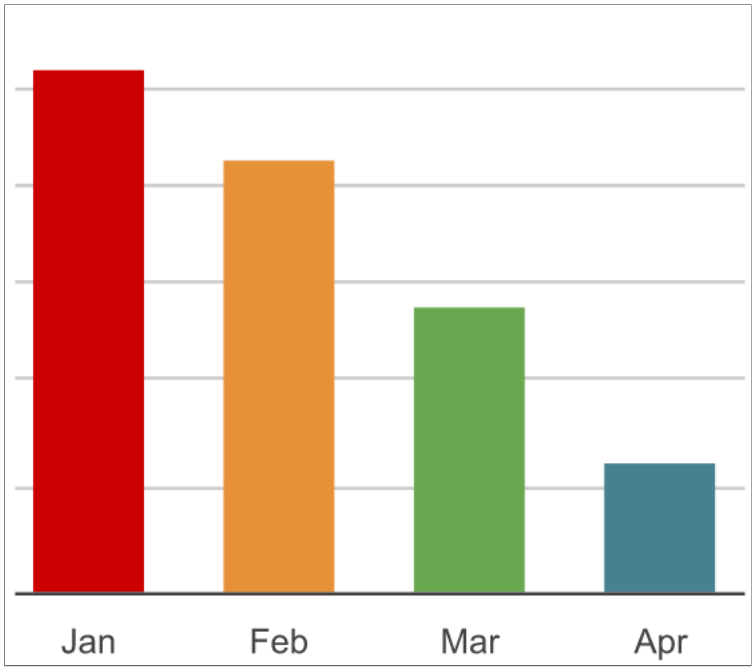
## Some tips Part 1

- don’t use 3-D bars



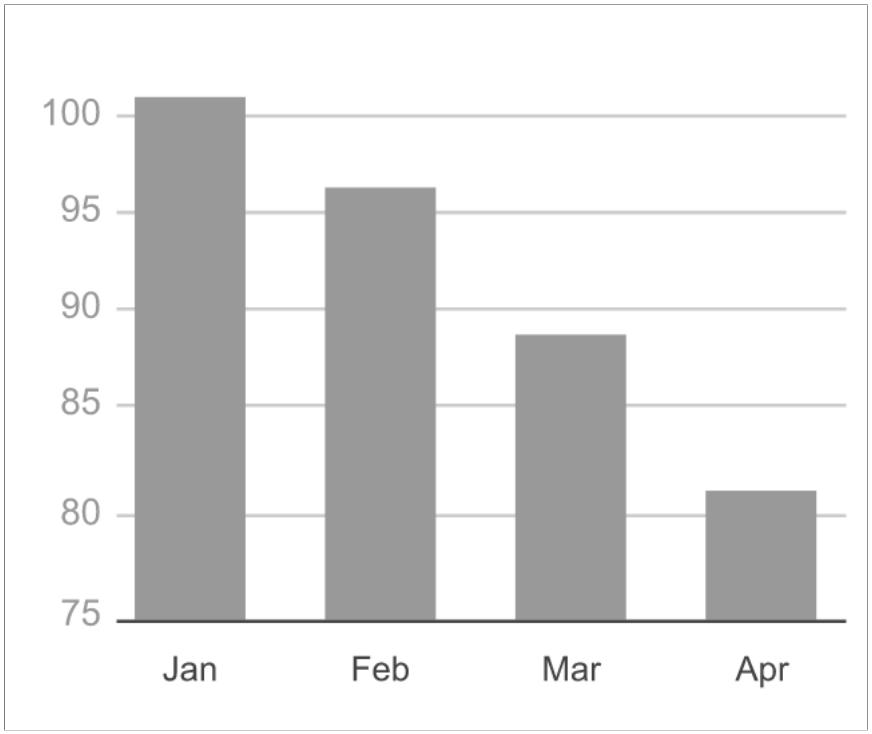
## Some tips Part 1

- don’t use 3-D bars
- don’t use multiple colors to represent the same kind of data



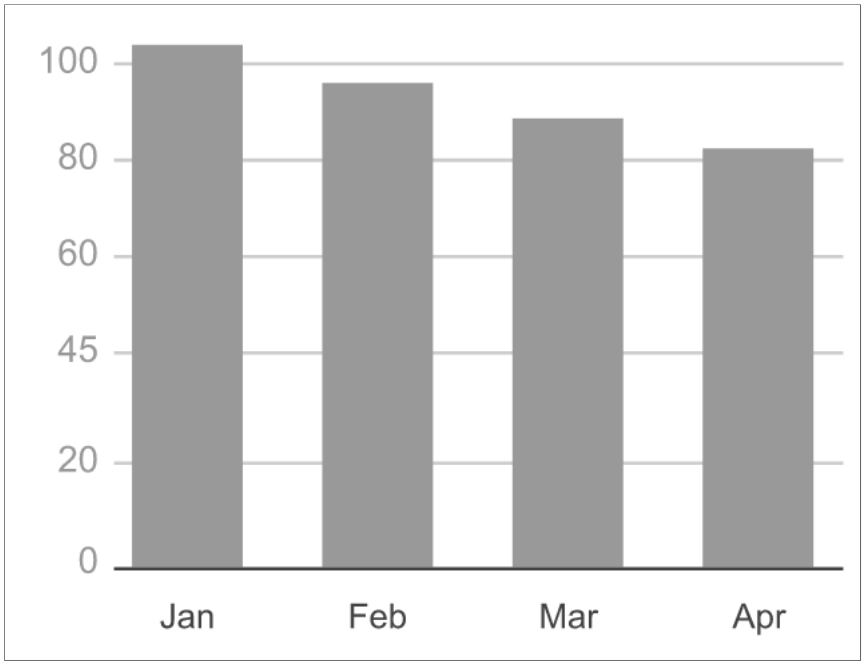
## Some tips Part 1

- don’t use 3-D bars
- don’t use multiple colors to represent the same kind of data
- don’t truncate baseline of bar charts



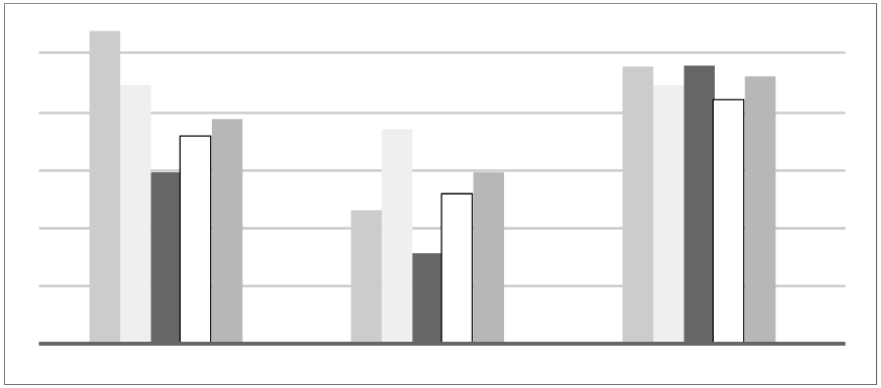
## Some tips Part 1

- don’t use 3-D bars
- don’t use multiple colors to represent the same kind of data
- don’t truncate baseline of bar charts



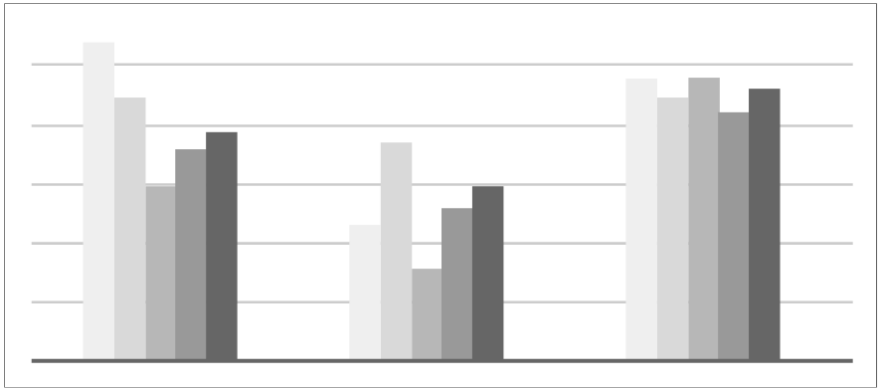
## Some tips Part 2

- no zebra pattern



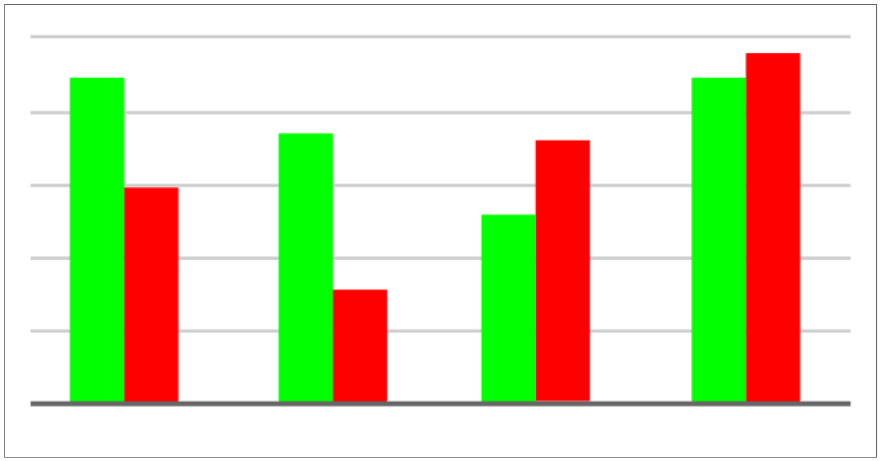
## Some tips Part 2

- no zebra pattern



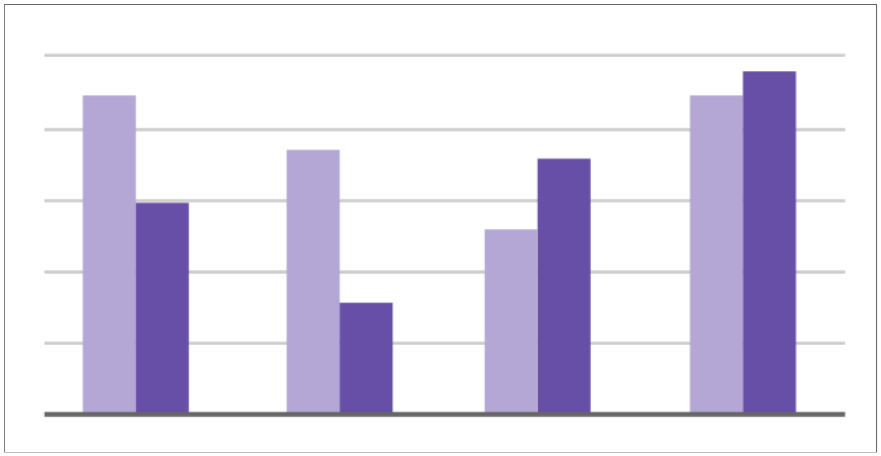
## Some tips Part 2

- no zebra pattern
- when working with two colors avoid complementary (opposite) colors



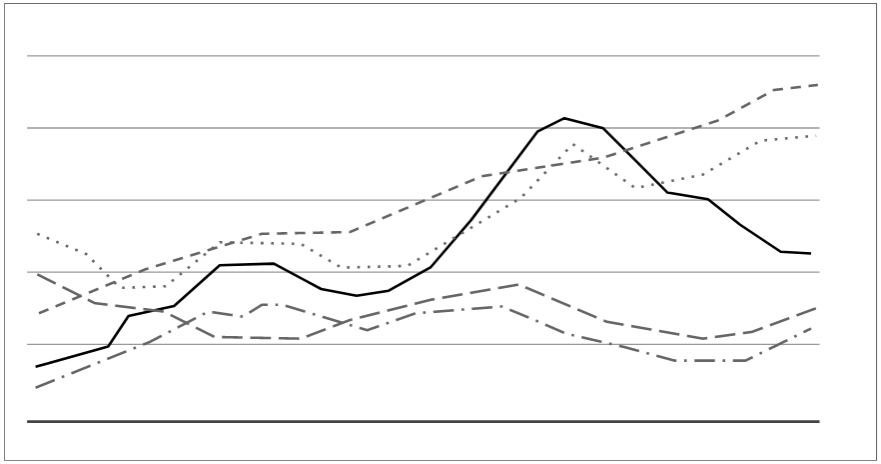
## Some tips Part 2

- no zebra pattern
- when working with two colors avoid complementary (opposite) colors



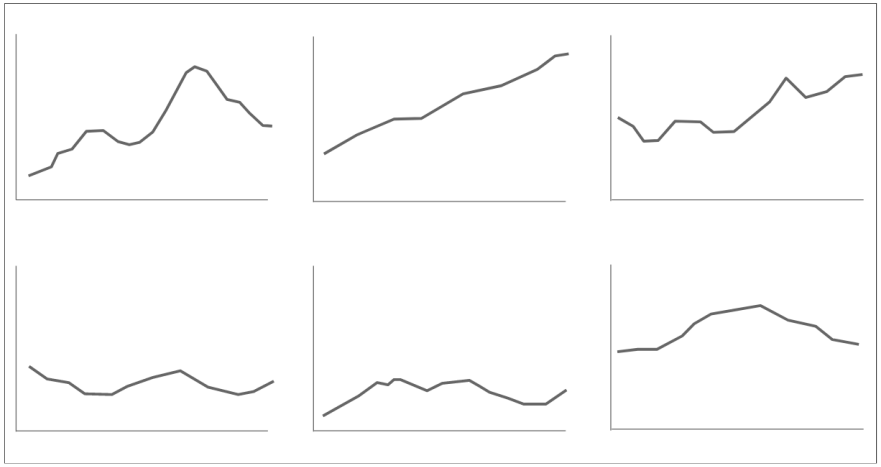
## Some tips Part 2

- no zebra pattern
- when working with two colors avoid complementary (opposite) colors
- avoid spaghetti lines



## Some tips Part 2

- no zebra pattern
- when working with two colors avoid complementary (opposite) colors
- avoid spaghetti lines



## Data inspection (distribution, normality)

Many statistical tests rely on *assumption of normality*

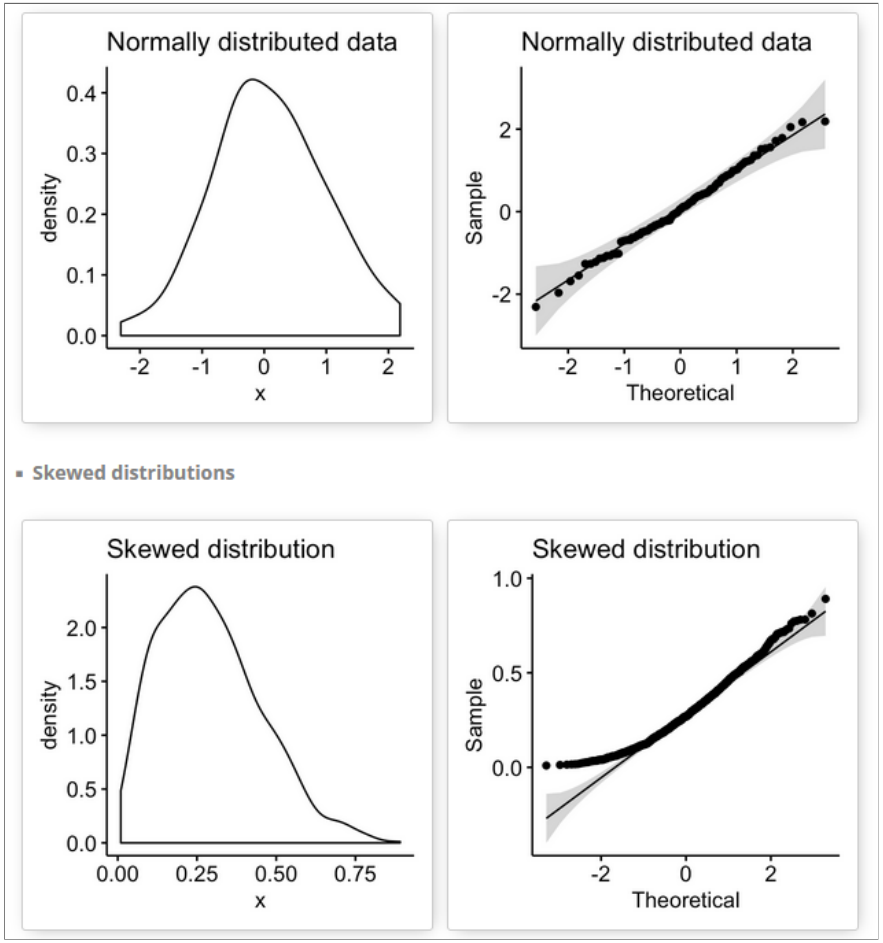
- one sample t-test
- two sample t-test
- ANOVA
- linear regression
- MRC
- Hierarchical linear modeling

If this assumption is violated the results of these tests become unreliable and we are unable to generalize our findings from the sample data to the overall population with confidence.

## How to check?

- visually
    1. histogram
    2. distribution
    3. Q-Q plot
    4. violin plot
- by formal statistical tests
    1. Shapiro-Wilk
    2. Kolmogorov-Smirnov (Lilliefors)
    3. Jarque-Bera
    4. Anderson-Darling
    5. Cramer-von Mises
    6. Pearson chi-square

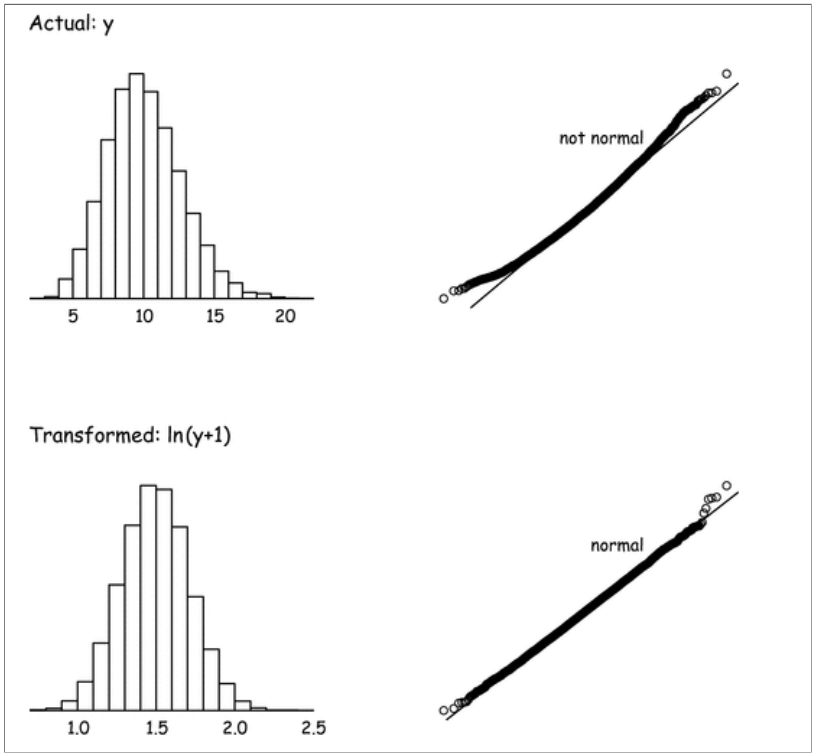
# Q-Q Plot


Normally distributed data
density
x
Normally distributed data
Sample
Theoretical
Skewed distributions
Skewed distribution
density
x
Skewed distribution
Sample
Theoretical

## What to do if we fail the test?

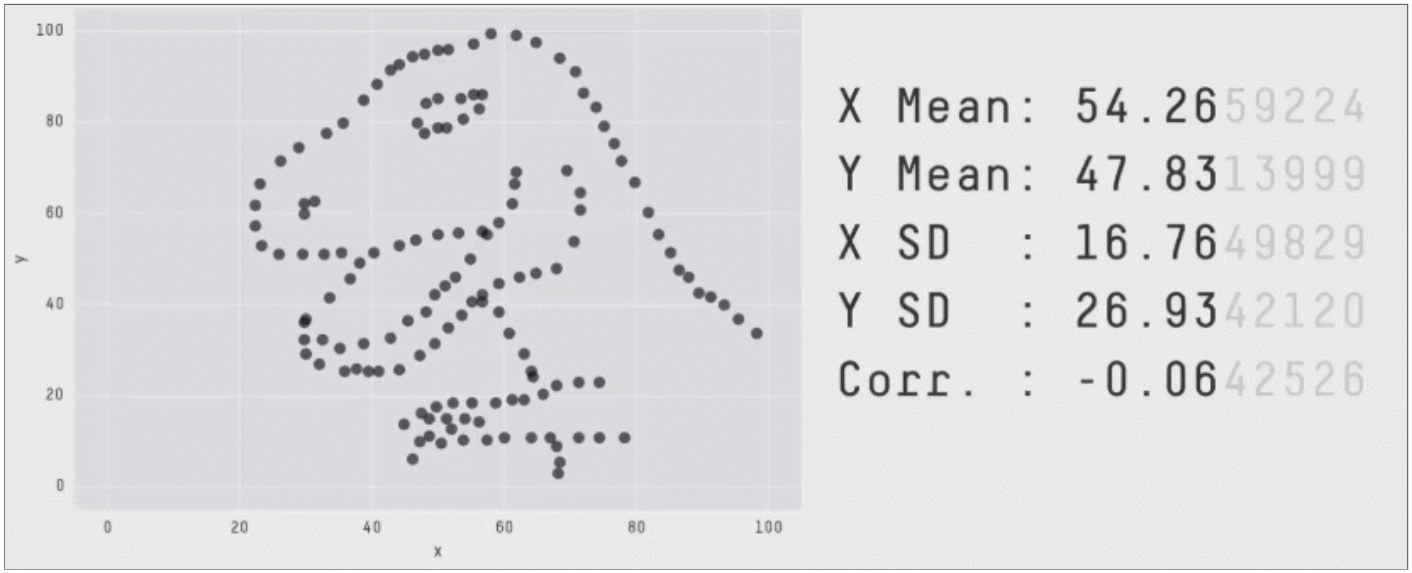
- transform data
  1. log transformation
  2. square root transformation
  3. cube root transformation
  4. Box-Cox transformation
- according to Central Limit Theorem, no matter what distribution is, the sampling distribution of mean tends to be normal if the sample is large enough ($n \geq 30$)
- use statistical tests for not normal distributions (non-parametric tests)

# Log transformation



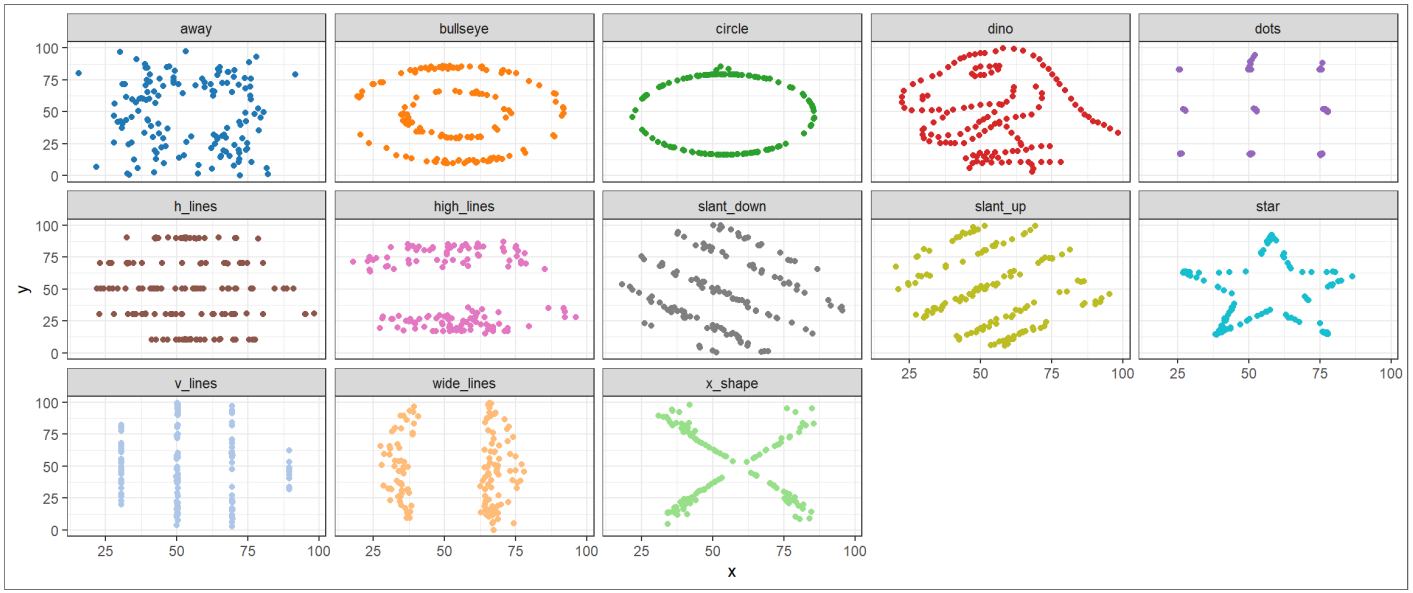
# Warning: DataSaurus

## Datasaurus


100
80
60
40
20
0
y
0
20
40
60
80
100
x
X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

# Datasaurus Dozen

```
datasaurus_dozen %>%
  ggplot(aes(x, y, color = dataset)) +
  geom_point(show.legend = FALSE) +
  facet_wrap(~dataset, ncol = 5)+theme_bw()+
 scale_color_d3("category20")
```

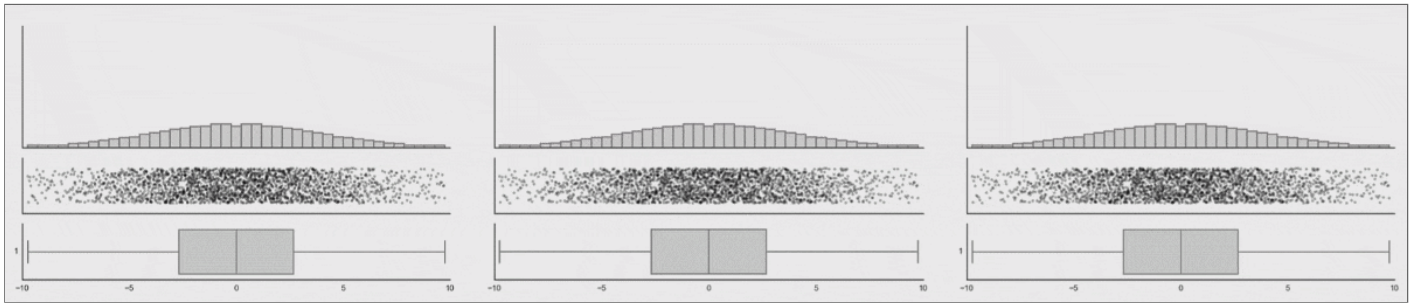


## Summary Statistics

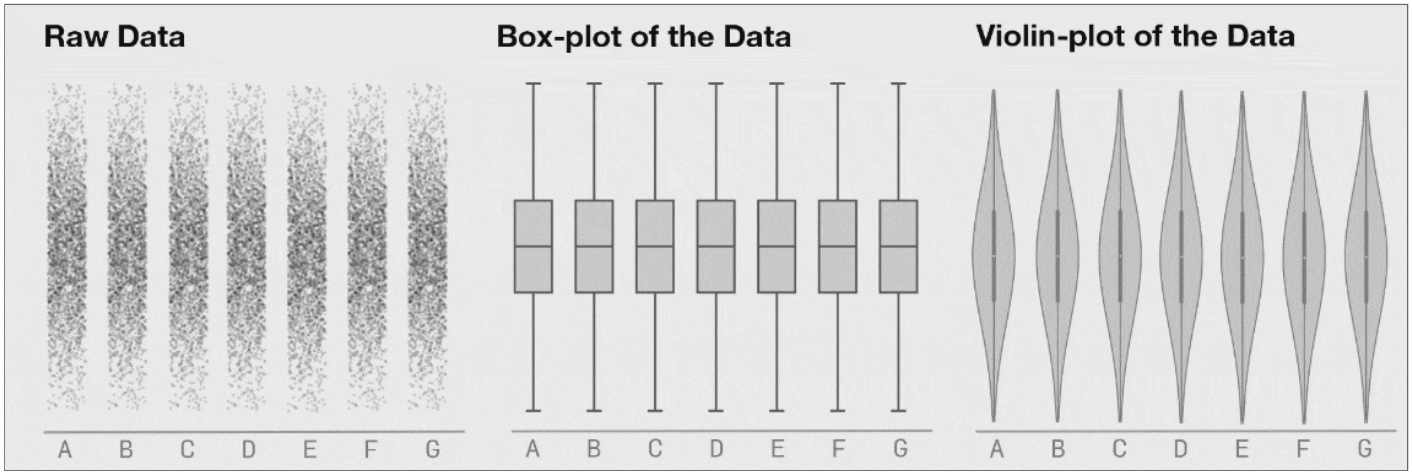
```
knitr::kable(
datasaurus_dozen %>%
  group_by(dataset) %>%
  summarise(across(c(x, y), list(mean = mean, sd = sd)),
    x_y_cor = cor(x, y)))
```

| dataset | x_mean | x_sd | y_mean | y_sd | x_y_cor |
|---|---|---|---|---|---|
| away | 54.26610 | 16.76983 | 47.83472 | 26.93974 | -0.0641284 |
| bullseye | 54.26873 | 16.76924 | 47.83082 | 26.93573 | -0.0685864 |
| circle | 54.26732 | 16.76001 | 47.83772 | 26.93004 | -0.0683434 |
| dino | 54.26327 | 16.76514 | 47.83225 | 26.93540 | -0.0644719 |
| dots | 54.26030 | 16.76774 | 47.83983 | 26.93019 | -0.0603414 |
| h_lines | 54.26144 | 16.76590 | 47.83025 | 26.93988 | -0.0617148 |
| high_lines | 54.26881 | 16.76670 | 47.83545 | 26.94000 | -0.0685042 |
| slant_down | 54.26785 | 16.76676 | 47.83590 | 26.93610 | -0.0689797 |
| slant_up | 54.26588 | 16.76885 | 47.83150 | 26.93861 | -0.0686092 |
| star | 54.26734 | 16.76896 | 47.83955 | 26.93027 | -0.0629611 |
| v_lines | 54.26993 | 16.76996 | 47.83699 | 26.93768 | -0.0694456 |
| wide_lines | 54.26692 | 16.77000 | 47.83160 | 26.93790 | -0.0665752 |
| x_shape | 54.26015 | 16.76996 | 47.83972 | 26.93000 | -0.0655833 |

# Boxplots

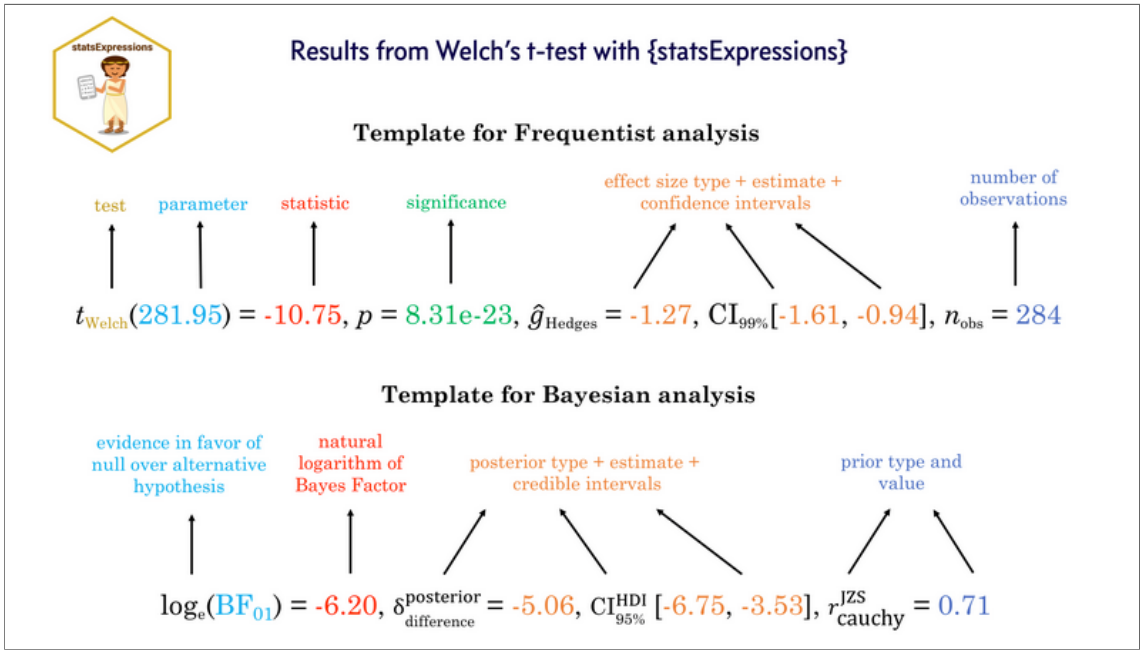


# Boxplots and Violin Plots

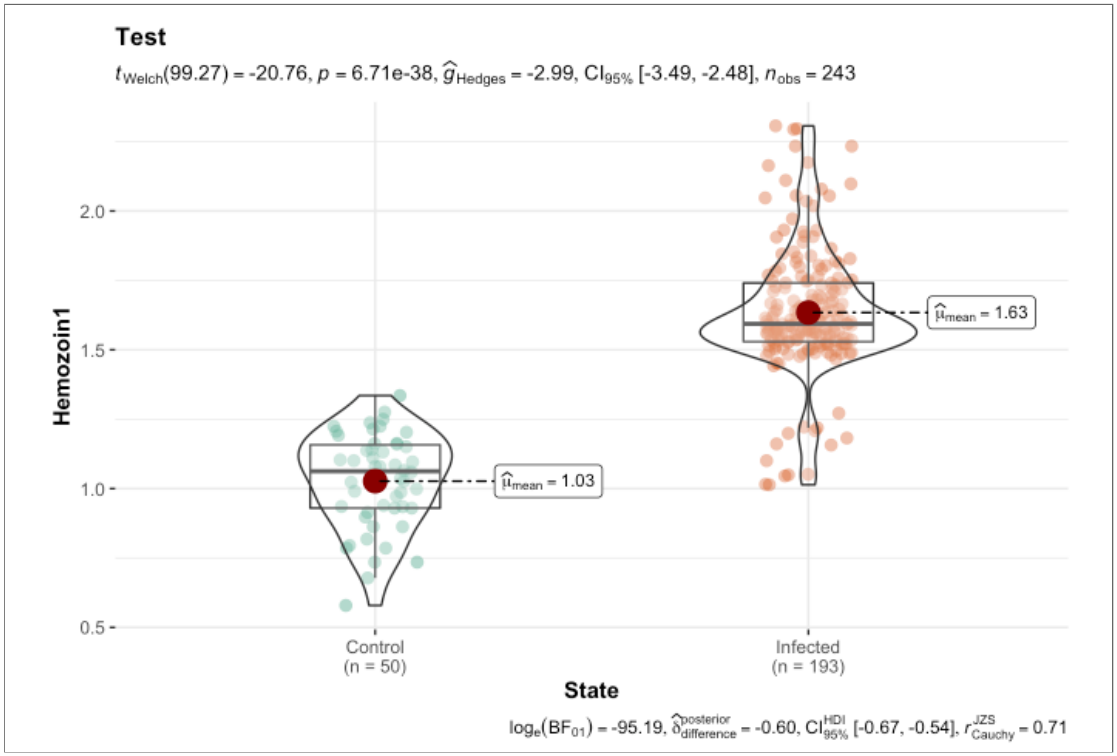


# Hypothesis testing

# Statistical Tests Reporting for Frequentist and Bayesians



# Reporting


Test
$t_{Welch}(99.27) = -20.76$, $p = 6.71e\text{-}38$, $\hat{g}_{Hedges} = -2.99$, $CI_{95\%}$ [-3.49, -2.48], $n_{obs} = 243$
Hemozoin1
2.0
1.5
1.0
0.5
$\hat{\mu}_{mean} = 1.03$
$\hat{\mu}_{mean} = 1.63$
Control
(n = 50)
Infected
(n = 193)
State
$\log_e(BF_{01}) = -95.19$, $\hat{\delta}^{posterior}_{difference} = -0.60$, $CI^{HDI}_{95\%}$ [-0.67, -0.54], $r^{JZS}_{Cauchy} = 0.71$

# Overview of multivariate statistical techniques

Unsupervised:

- PCA
- k-Mean Clustering
- HCA

Supervised:

- Naive Bayes Classifier
- k-Nearest Neighbours
- Linear Regression
- Logistic Regression
- Tree-Based Models
- Artificial Neural Networks
- Support Vector Machine

## Meta analysis and model validation

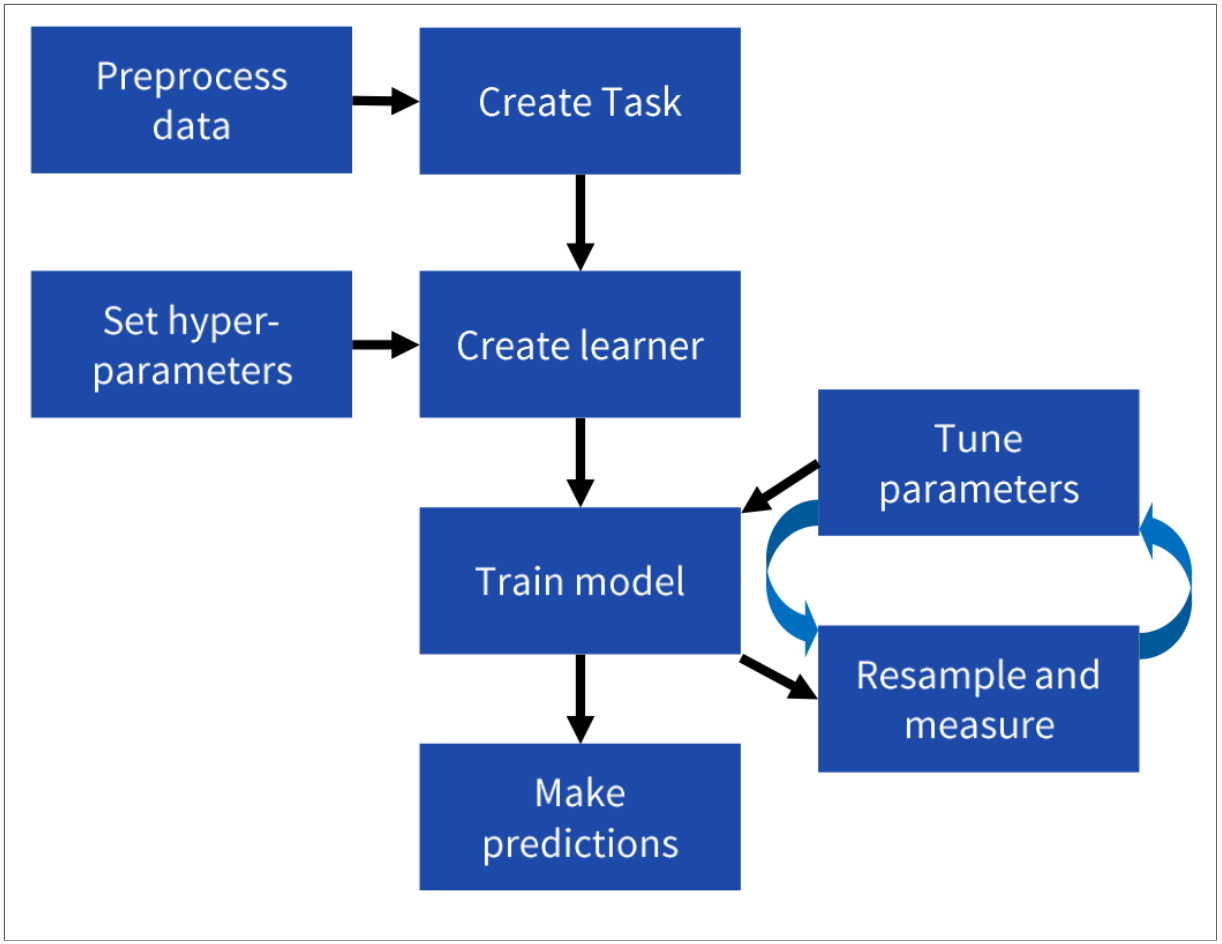
Meta Analysis:

- Regularisation
- Boosting
- Bagging
- Prunning
- Random Forest

Model Validation:

- Performance Metrics
- Bias-Variance Trade-off
- Cross-validation
- Learning Curves

## Machine Learning Workflow



### Stages

- pre-processing data,
- creating a task and making a learner,
- setting hyperparameters, train a model and predict,
- measuring performance,
- resampling a learner and tuning hyperparameters,

## Experiment

## Data preprocessing

- data quality check
- baseline correction
- outlier removal
- normalization/scaling

## Data Evaluation

- dimension reduction
- clustering, grouping
- visualization
- reporting

## Results reporting

- reports, presentations
- journal articles, chapters, books
- blog, website

## Resources

**https://www.dataquest.io/data-science-resources/** Data Science Resources
**https://r-charts.com/** R Charts
**https://r-graph-gallery.com/** R Graph Gallery

## Take home message(s)

- scientific method is still valid approach in the research so:
  1. hypothise
  2. measure
  3. analyse
  4. repeat!
- take care of your data distribution and adequate statistical tests
- visualization of data is of utmost importance
- do not trust boxplots or summary statistics (mean, SD etc)

## THANK YOU FOR YOUR ATTENTION