VISUALISING DATA: GOOD PRACTICE AND WORKFLOW

National Workshop

Accra, Ghana

Outline

1. Properties of good graphs
2. Statistical workflow
3. A brief introduction to R

1-1. Properties of good graphs: how to present

Aims

- Note: good graphs are self-explanatory!
  - The key to understand a graphs should not be hidden somewhere in the text!
- Often, the optimal presentation of data is not “standard”.
- There are no “recipes” how to present data.
- We have to use our own imagination.
- Still, some examples might help.

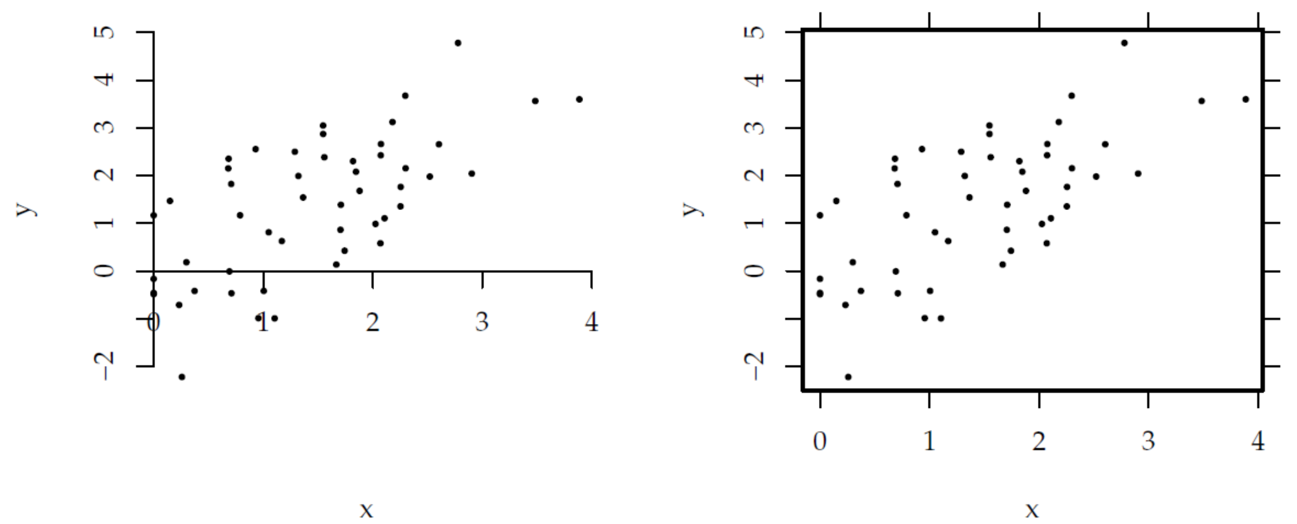
What can good graphs achive?

- A graph can…
  - … create trust
  - … motivate a question,
  - … summarise conclusions of the paper.
- A graph must be very good:
  - Some readers look only at figures and graphs.
  - Each graph should tell a story.
  - Among the many ways to present our data and our results, we have to chose the best way.

1-1-1. Axes

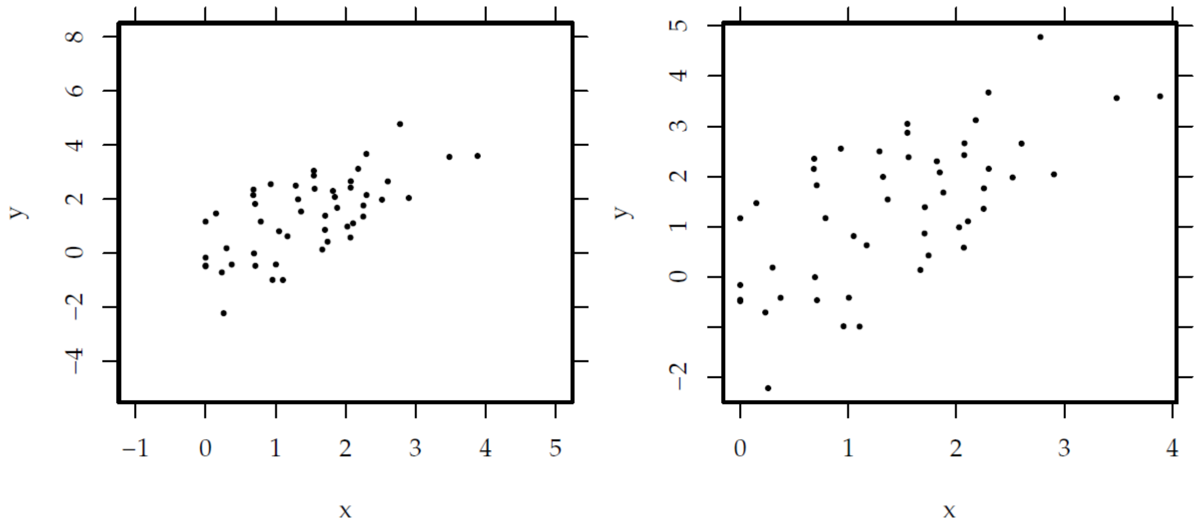
Frames

- Labels and tick marks separated from the data
- Ticks on the opposite axis can help



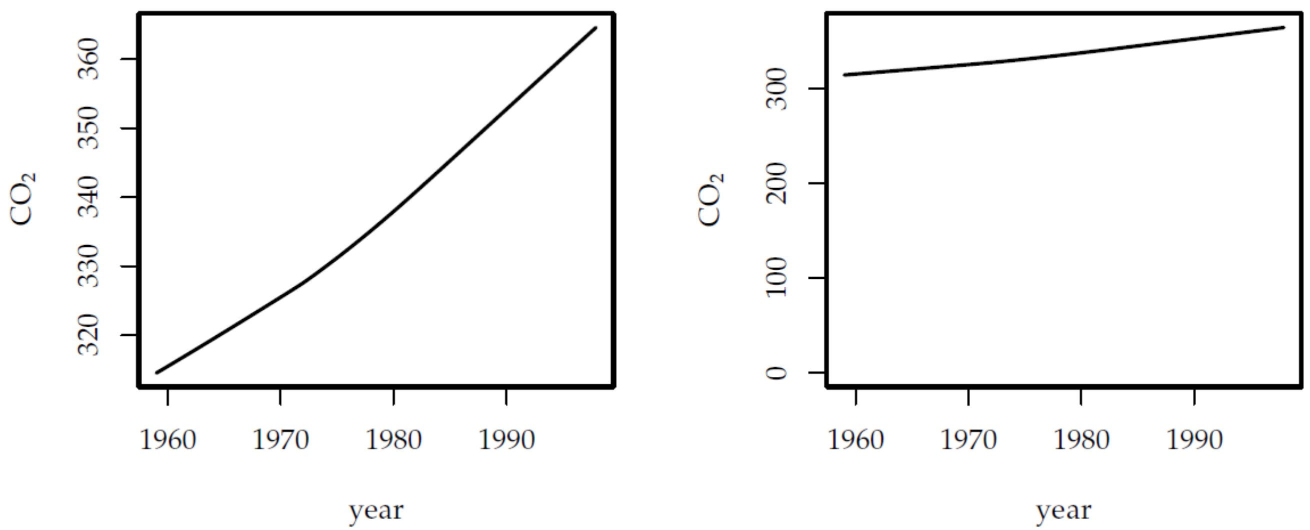
Ranges

- Ranges are chosen such that
  - all data is included,
  - space is used in an efficient way.



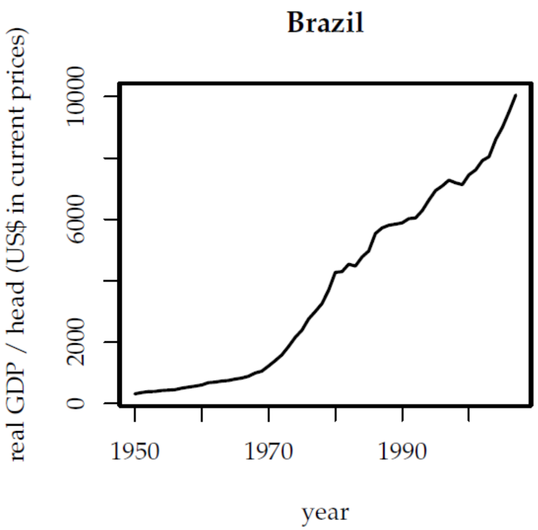
Ranges that include zero

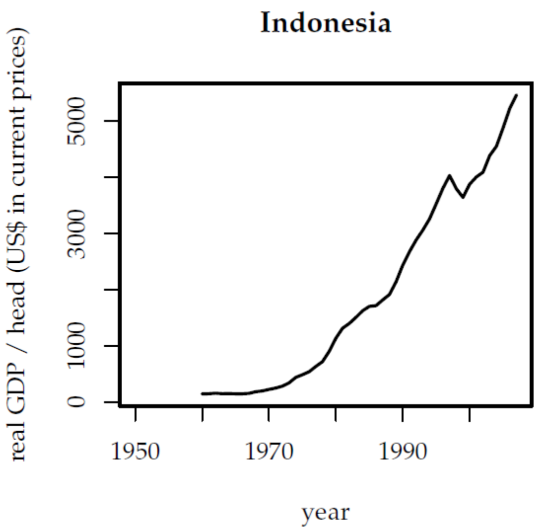
- It is often helpful to include zero, but since this might waste space it is not an absolute necessity



Comparable scales (1)

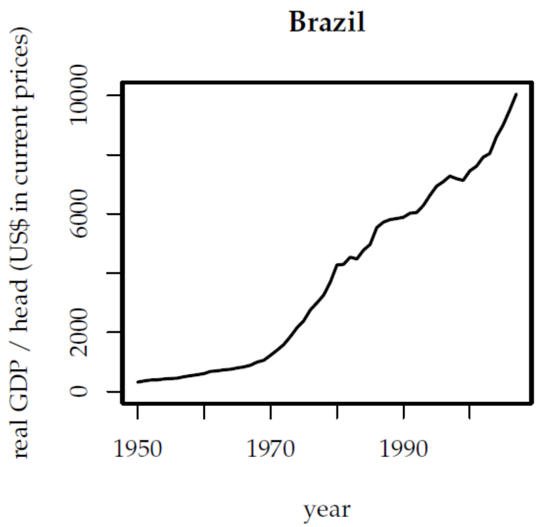
- The two diagrams use different scales.
- This makes them difficult to compare.

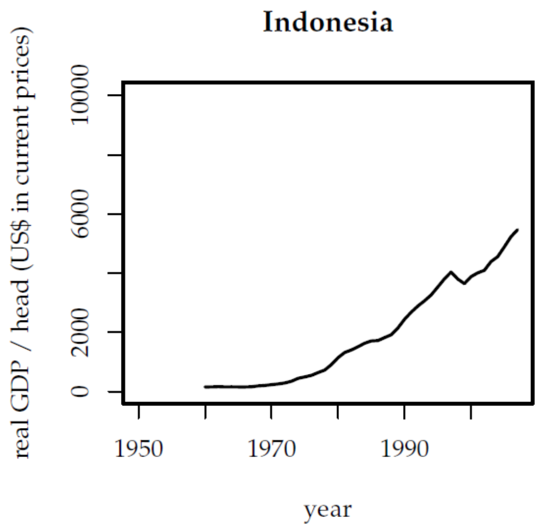




Comparable scales (2)

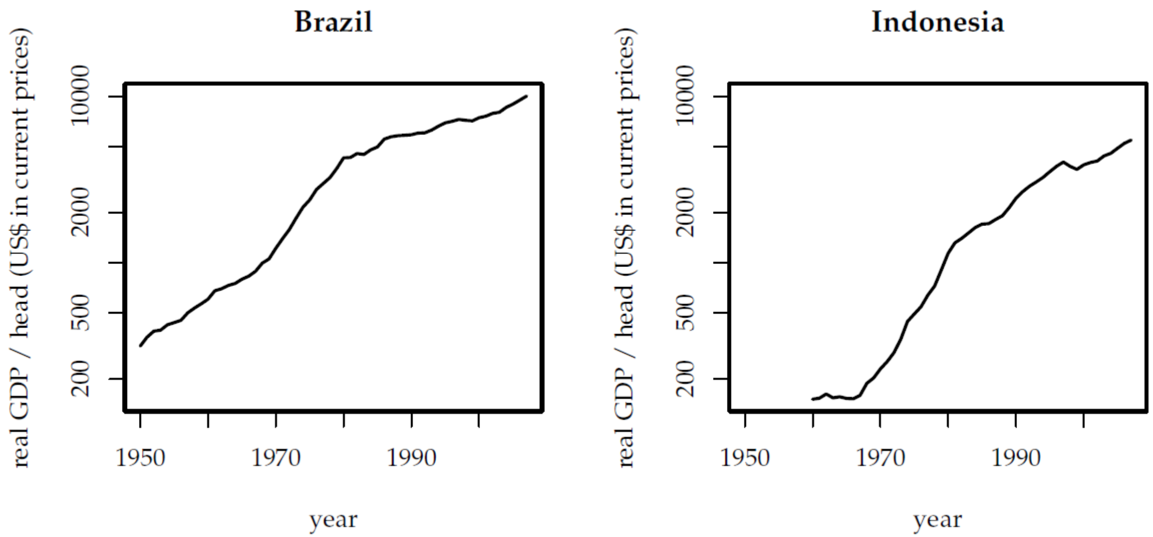
- The two diagrams use the same scale.
- Now we see that GDP is larger in Brazil.





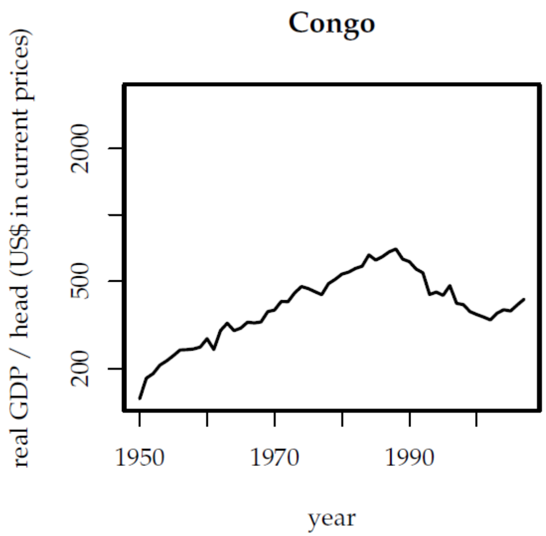
Comparable scales (3)

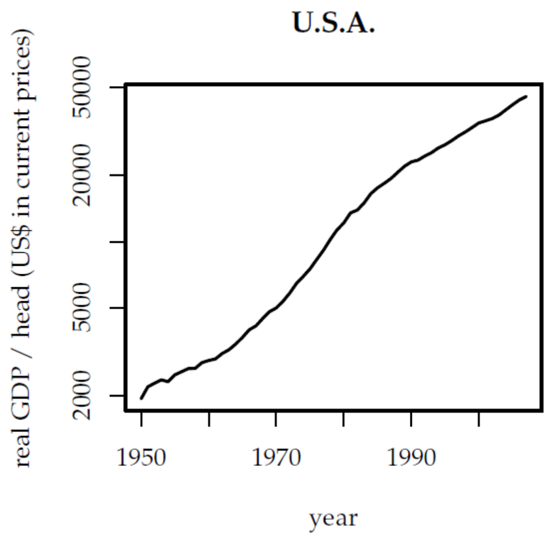
- Using a logarithmic scale allows to compare relative growth.



Comparable scales (4)

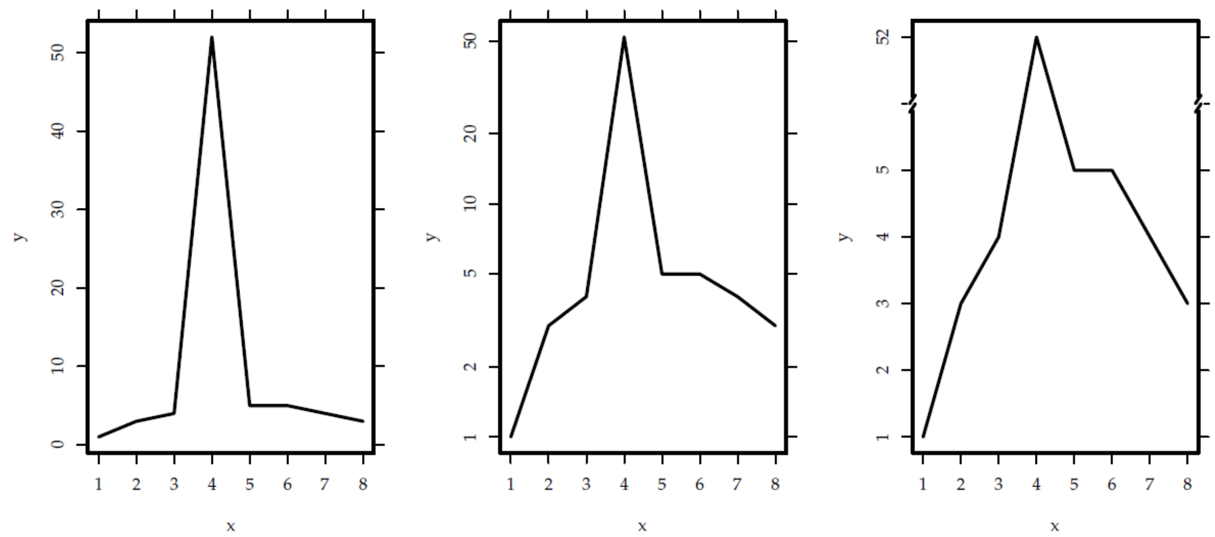
- It is not always necessary to use the same axes.





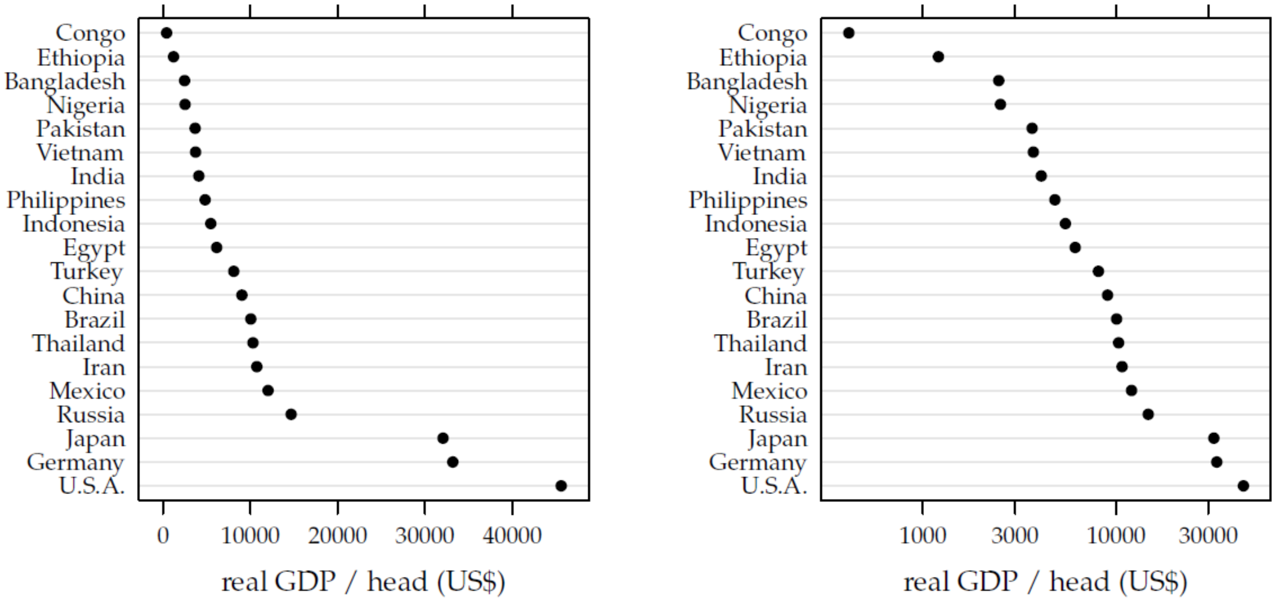
Breaks

- All three graphs try to show the same data.
  - Linear scale, logarithmic scale, and breaking the axis



Logarithmic scale

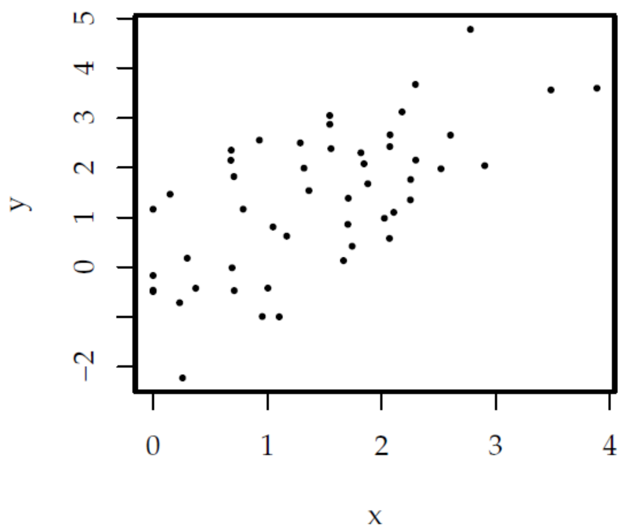
- The graph on the left shows GDP on a linear scale, the one on the right uses a logarithmic scale.

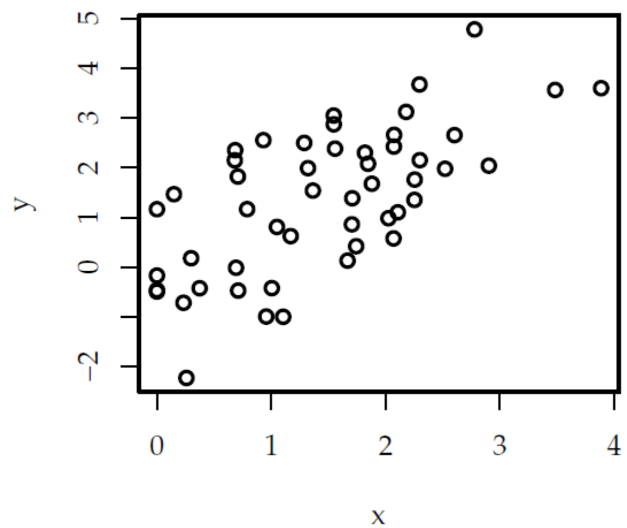


1-1-2. Points

Points (1)

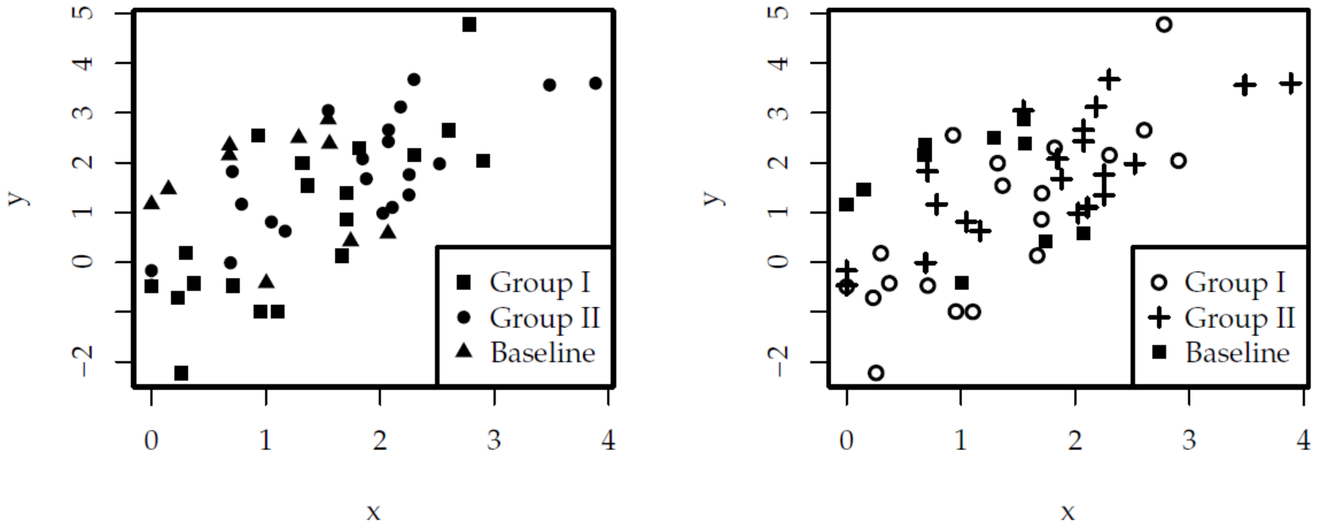
- The graph on the left uses fairly small points, the one on the right uses larger points.





Points (2)

- Plotting symbols should be easy to distinguish



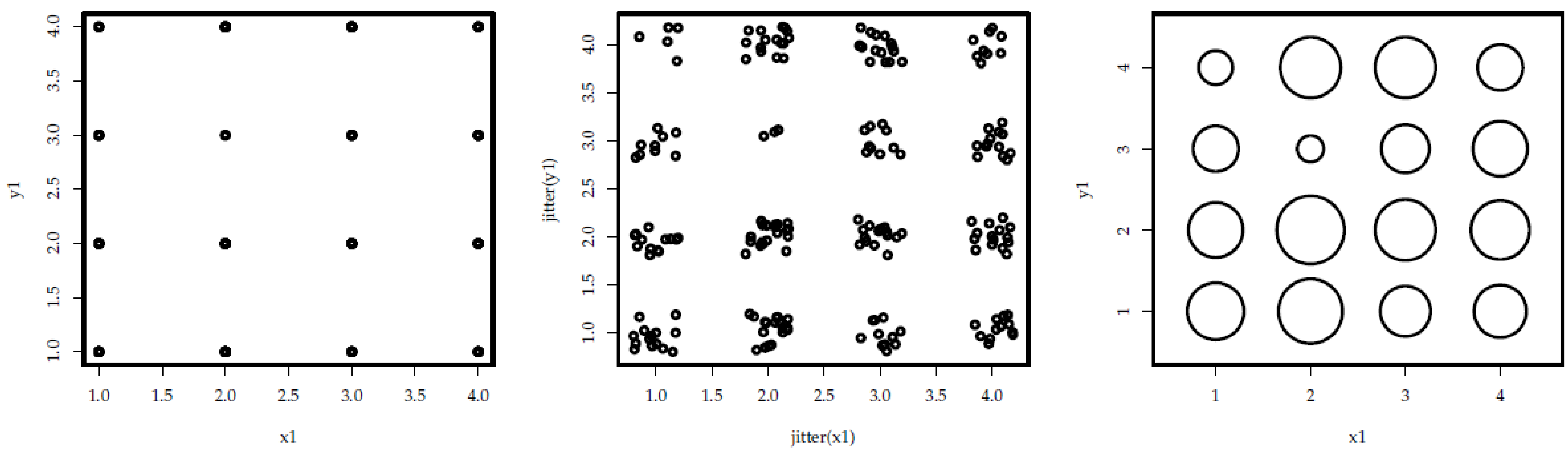
Kröse‘s experiment

- Participants see patterns of symbols for 80 ms. They have to identify presence or absence of a given symbol.

symbols	% recognised
$+\circ$	100.0
$+\square$	88.1
$L+$	68.6
$\Delta\downarrow$	52.3
$+T$	37.6
$+X$	30.3
TL	30.6

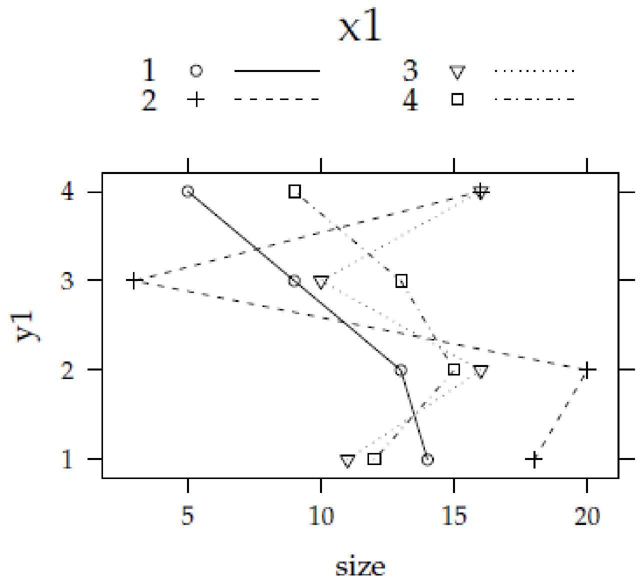
Nominal data and points (1)

- Sometimes, in particular with nominal data, we want to show the same obversation several times.
- Distinguish observations using
  - Jitter or frequency represented by symbols



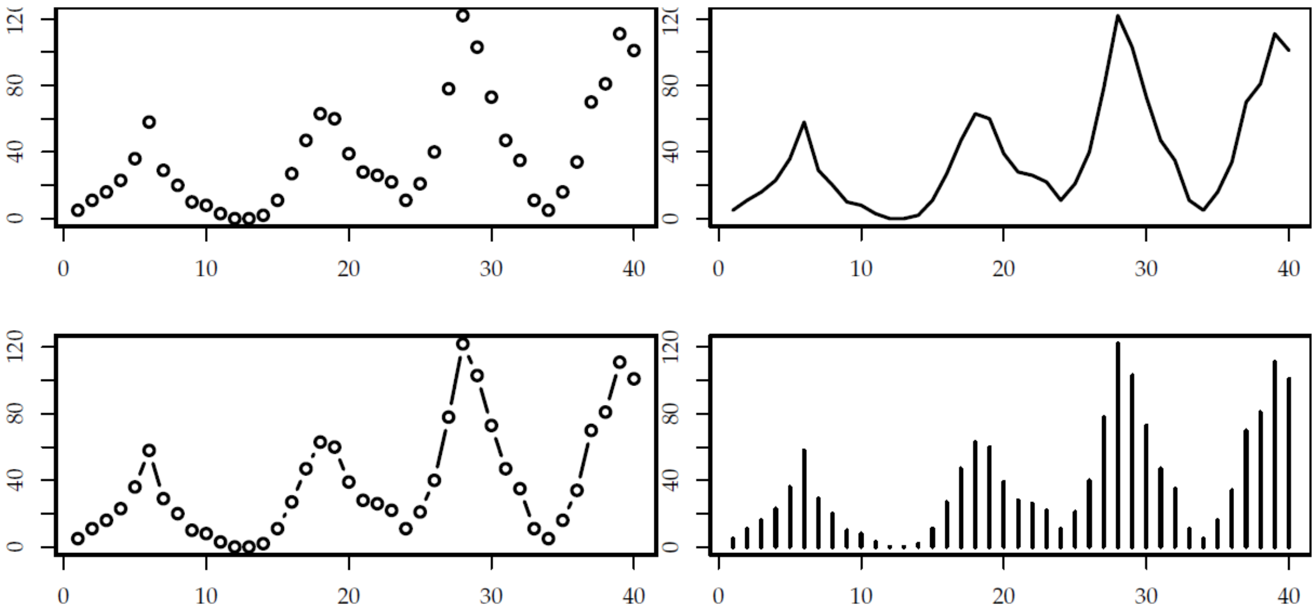
Nominal data and points (2)

- If we have a small number of categories (at least in one dimension), a dot plot might be better



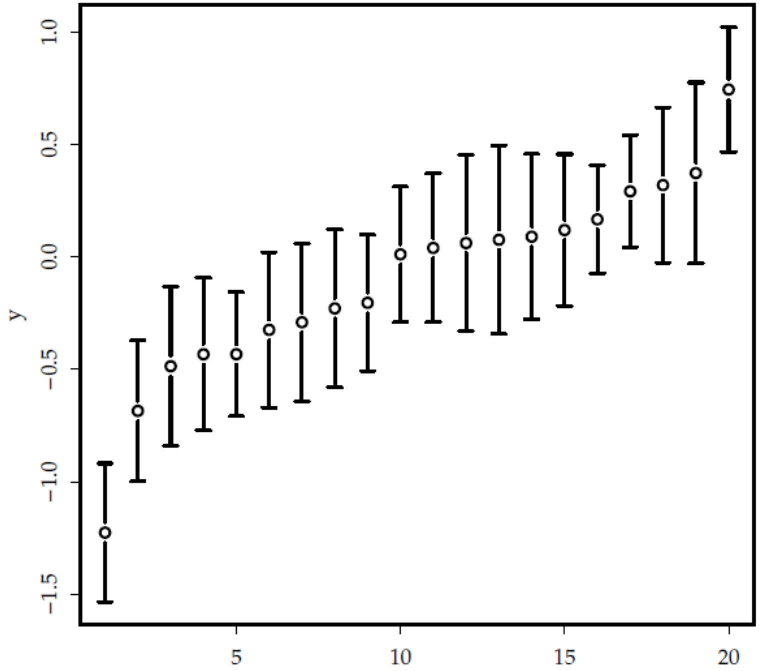
Points and time series

- Time trends are easier to see with lines.
- Lines alone make it impossible to find out when the measurements were taken.



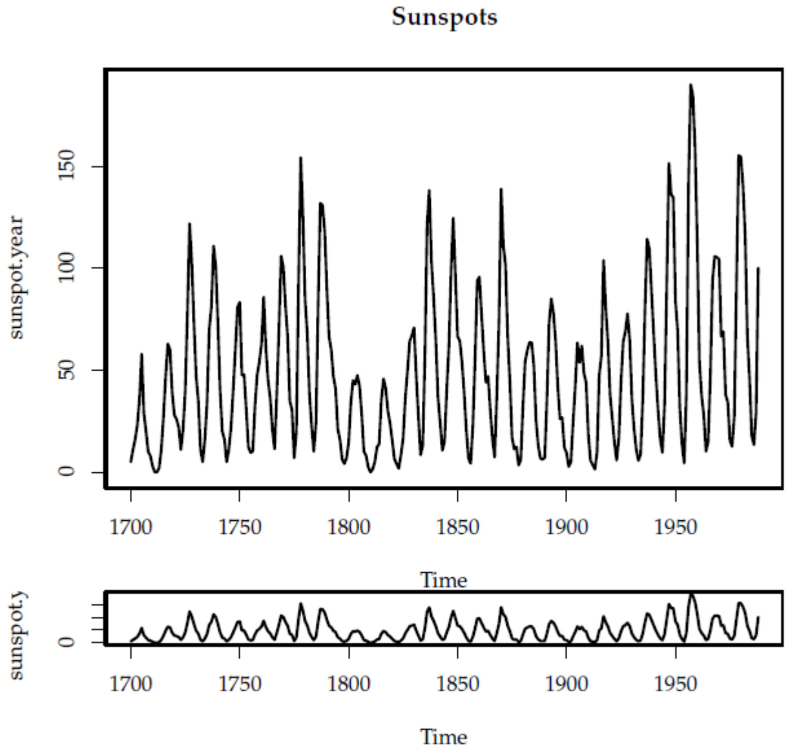
Error bars

- Explain clearly in figure what quantity is shown.
- Boxplots often more informative than error bars.



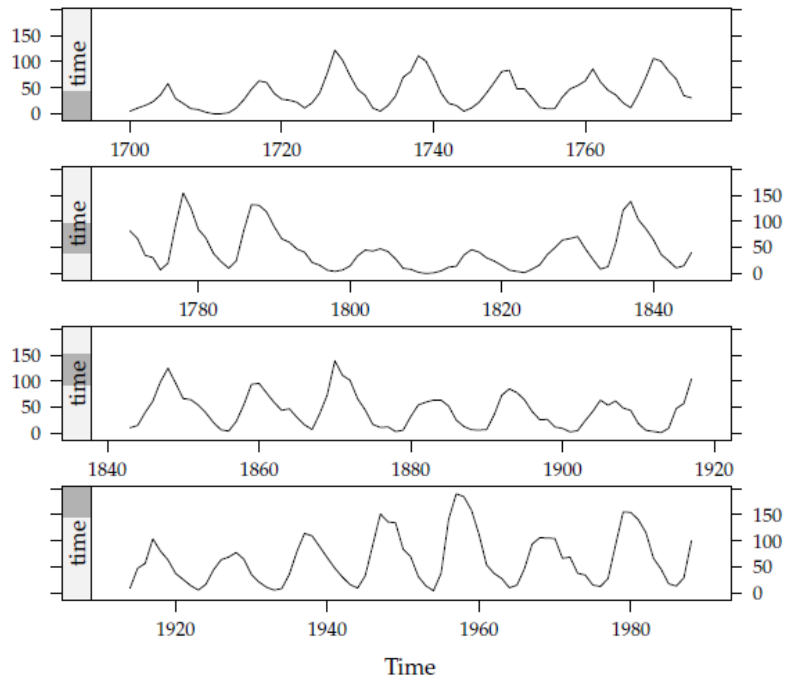
Aspect ratio (1)

- Less can be more



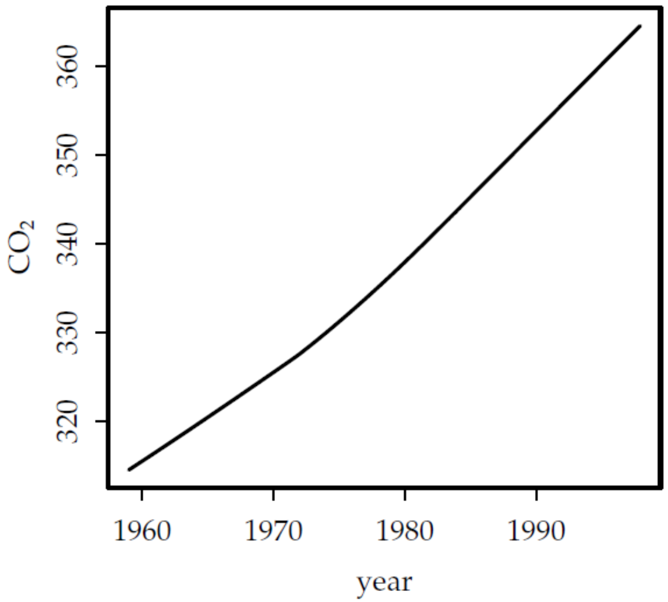
Aspect ratio (2)

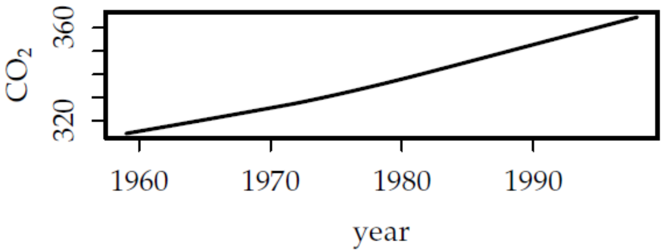
- If we feel that the lower graph is too flat then we can 'cut-and-stack' it



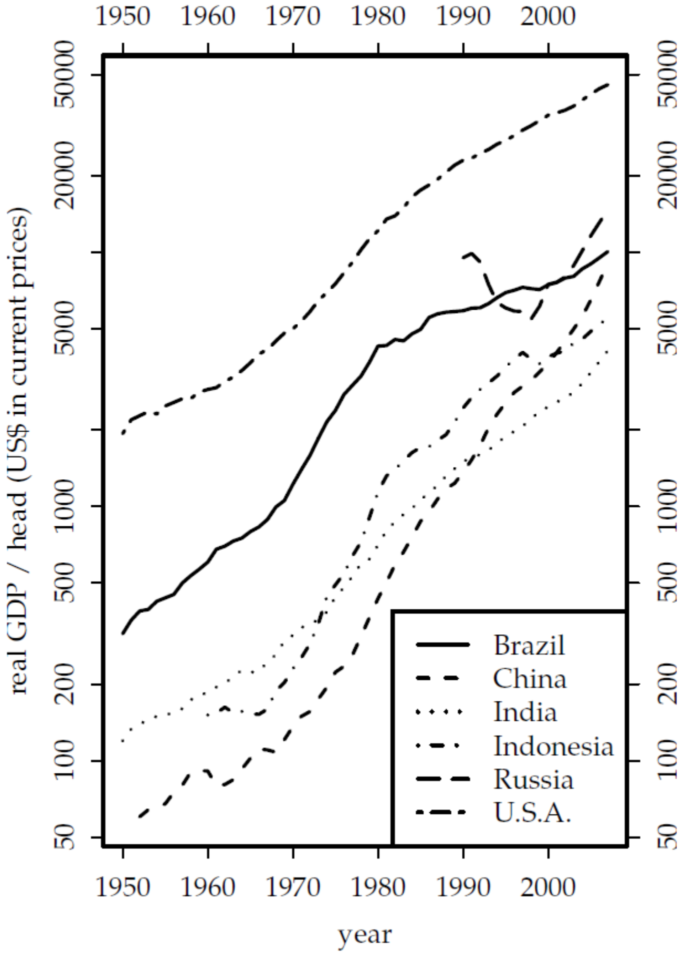
Aspect ratio (3)

- The graph on the left has a slope of about 45%.
- This makes it easier to see the convexity of the curve.





Aspect ratio (4)

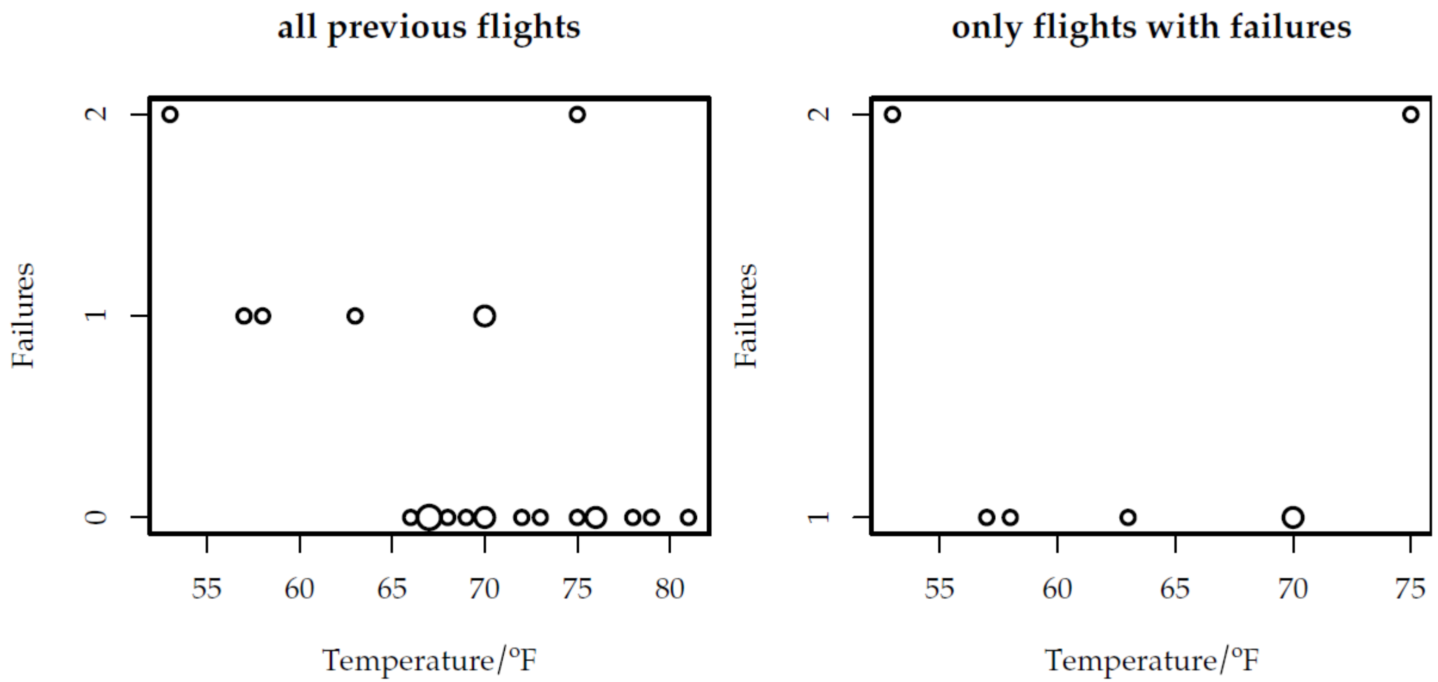


- Lines have a slope of about 45%.
- This makes it easier to compare the different slopes.

1-2. Properties of good graphs: what to present

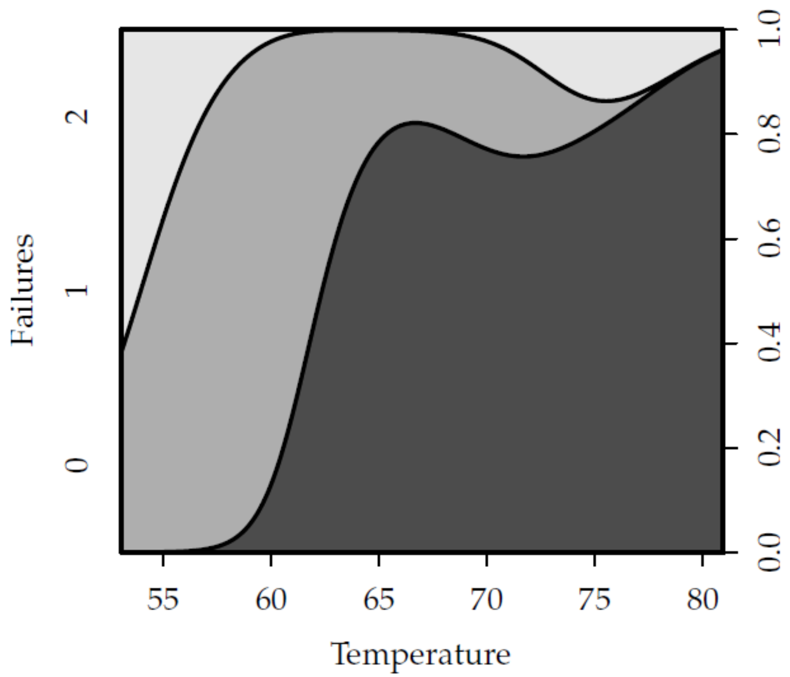
Don‘t discard parts of your data (1)

- Previous to the crash of the space shuttle Challenger, managers only inspected failures.



Don‘t discard parts of your data (2)

- An alternative way to present this data is a conditional density plot



2. Statistical workflow

A worked example

- check folder: **\project\permanent\3_R**

3. A brief introduction to R

Learning R, in R.

```
## install swirl package
> install.packages("swirl")
> library(swirl)
> install_from_swirl("R Programming")

## start swirl
> swirl()
| Please choose a course -- select "7: R Programming"
> 7
| Please choose a lesson -- work through lessons 1, 2 and 12
> 1
## you can exit any time by pressing the 'Esc' key
```

Sign up for the R programming course today!
Check: https://www.coursera.org/course/rprog

Data import

- read.csv function
  - dat = read.csv(file='D:/project/data.csv', [options])
- options
  - stringsAsFactors = FALSE # avoid conversion to factors
  - sep = "," # specify the field separator

Data types

```
## numeric
NUM <- 1:10
str(num)

## integer
INT <- rnorm(10)

## logical
LOG <- ifelse(NUM>5, TRUE, FALSE)

## character
CHR <- letters
paste(CHR)

## factor
FAC <- as.factor(CHR)
```

Data type conversion

- Conversion between data types

	numeric	character	factor
numeric	--	as.character(NUM)	as.factor(NUM)
character	as.numeric(as.factor(CHR))	--	as.factor(CHR)
factor	as.numeric(FAC)	as.character(FAC)	--

Related resources for R

- Data import
  - Work through
    - ats.ucla.edu/stat/r/faq/inputdata_R.htm
  - Also see
    - statmethods.net/input/importingdata.html
    - r-tutor.com/r-introduction/data-frame/data-import
- Data type conversion
  - Work through
    - cookbook-r.com/Manipulating_data/Converting_between_vector_types/
  - Also see
    - statmethods.net/management/typeconversion.html