Data Visualisation Using ggplot2

Dr Rafael de Andrade Moral
Associate Professor of Statistics, Maynooth University

rafael.deandrademoral@mu.ie
https://rafamoral.github.io

Outline

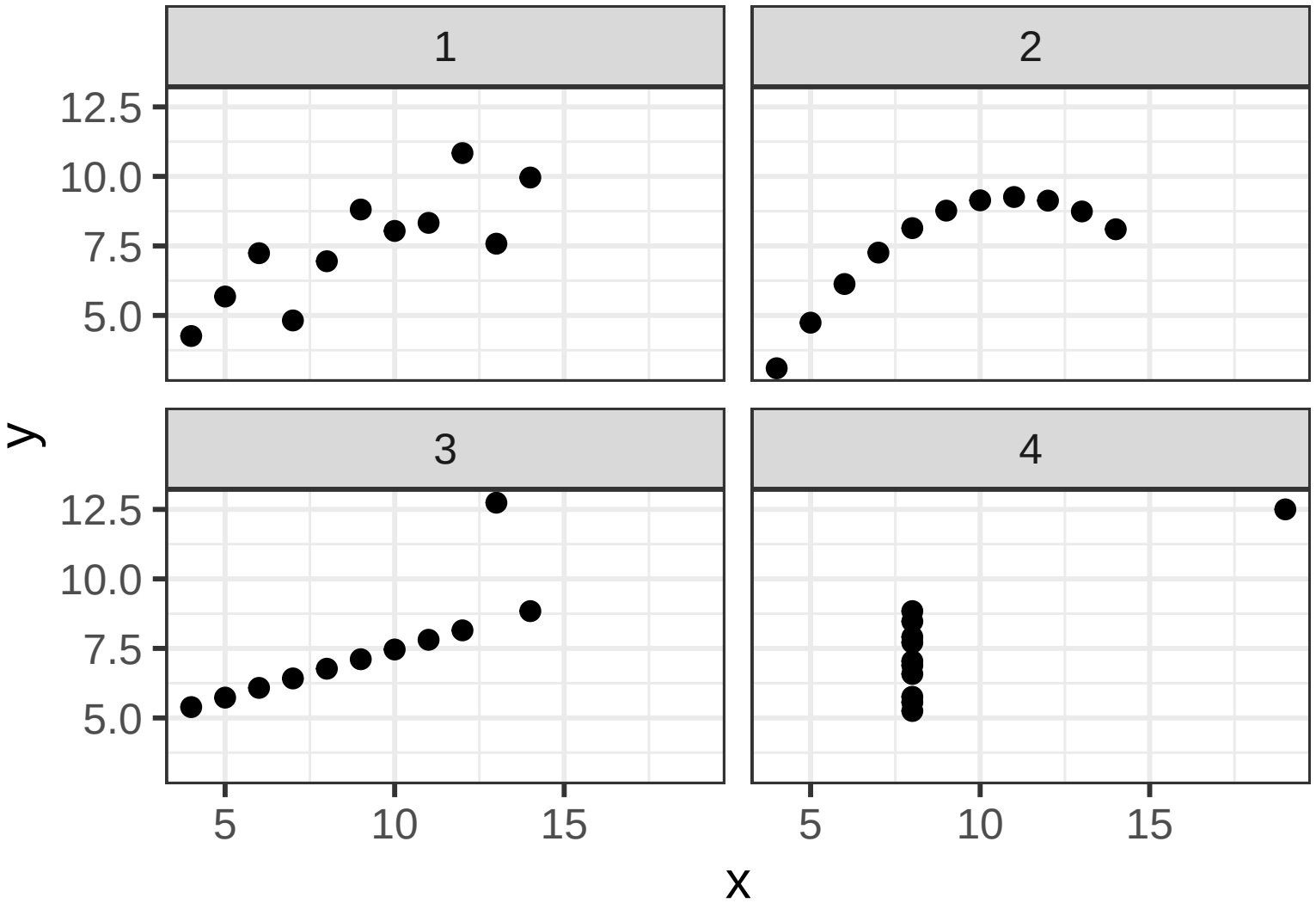
- What is data visualisation
- Introducing `ggplot2`
- Visualising univariate data
- Scatterplots
- Other plot types
- Fine control
- Plots for publications and presentations

Data Visualisation

- *The ability of the human eye to find patterns in scatters of points is a strong reason to use graphical techniques* (A. C. Atkinson, 1985)
- Rather than being a means to add some eye-candy or ornamentation to otherwise dull reports or slides, the purpose of visualisation is to allow us explore data and find patterns that would easily be missed were we to rely only on numerical summary statistics.
- A classic illustration of this is Anscombe's quartet (Anscombe, 1973):

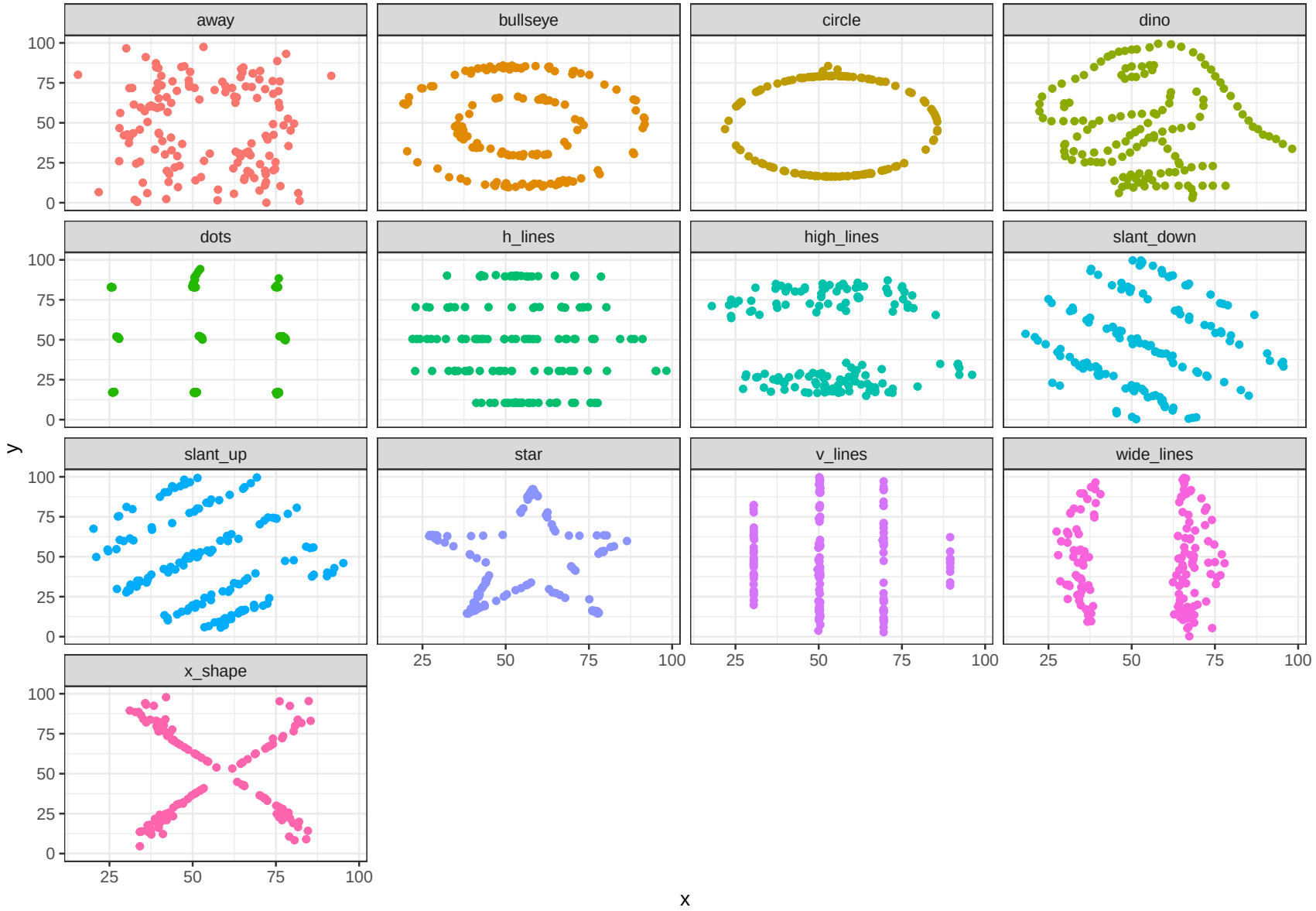
set	mean(x)	mean(y)	sd(x)	sd(y)	cor(x,y)
1	9	7.5	3.32	2.03	0.82
2	9	7.5	3.32	2.03	0.82
3	9	7.5	3.32	2.03	0.82
4	9	7.5	3.32	2.03	0.82

Anscombe's Quartet



The DataSaurus Dozen

dataset	mean(x)	mean(y)	sd(x)	sd(y)	cor(x,y)
away	54.27	47.83	16.77	26.94	-0.06
bullseye	54.27	47.83	16.77	26.94	-0.07
circle	54.27	47.84	16.76	26.93	-0.07
dino	54.26	47.83	16.77	26.94	-0.06
dots	54.26	47.84	16.77	26.93	-0.06
h_lines	54.26	47.83	16.77	26.94	-0.06
high_lines	54.27	47.84	16.77	26.94	-0.07
slant_down	54.27	47.84	16.77	26.94	-0.07
slant_up	54.27	47.83	16.77	26.94	-0.07
star	54.27	47.84	16.77	26.93	-0.06
v_lines	54.27	47.84	16.77	26.94	-0.07
wide_lines	54.27	47.83	16.77	26.94	-0.07
x_shape	54.26	47.84	16.77	26.93	-0.07


away
bullseye
circle
dino
dots
h_lines
high_lines
slant_down
slant_up
star
v_lines
wide_lines
x_shape
y
x
0
25
50
75
100

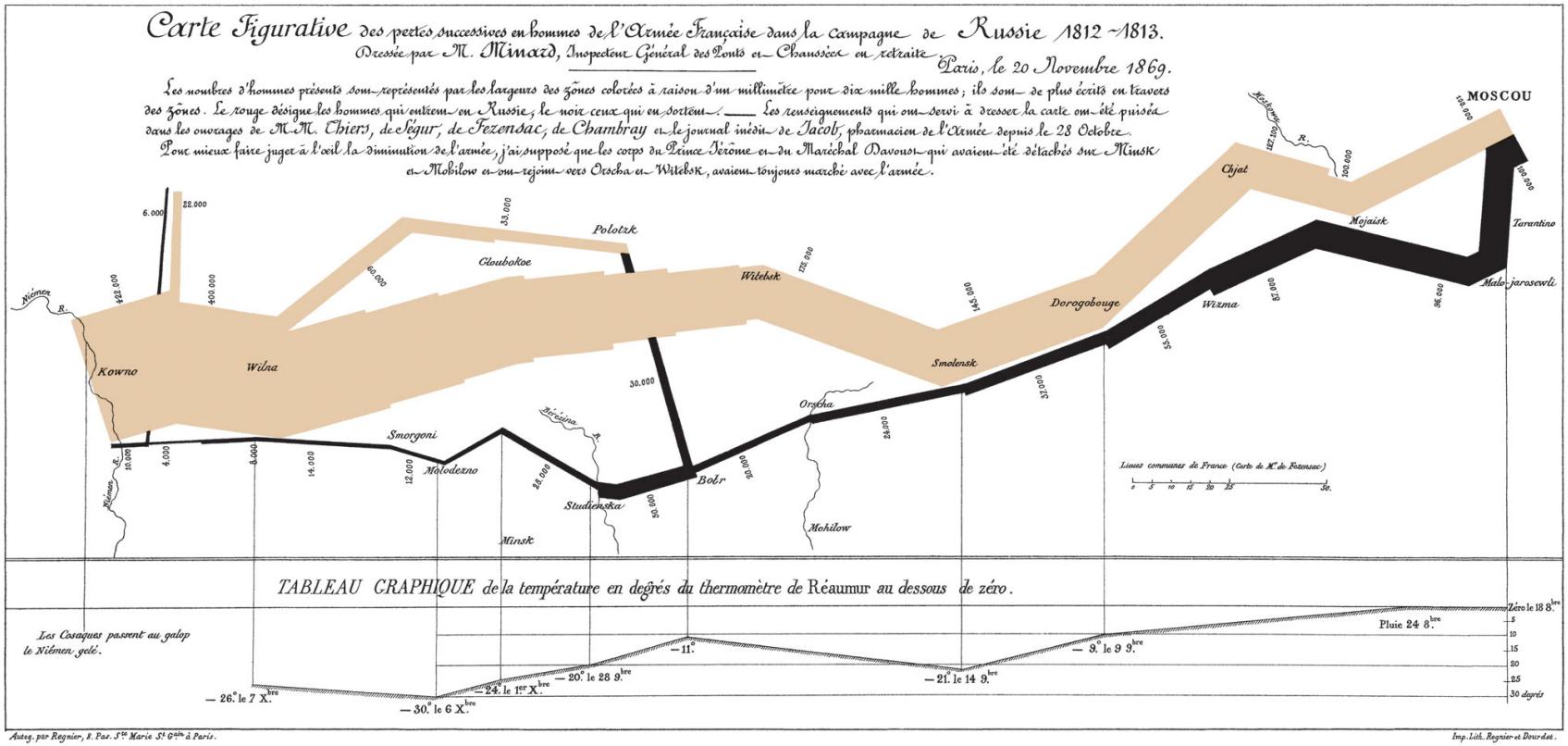
The Aim of Data Visualisation

- A key characteristic of data visualisation, therefore, is that *it forces us to notice what we never expected to see* (Tukey, 1977)
- In other words, data visualisation is not simply a means to graphically illustrate what we already know, but to reveal patterns and structures in the data
- Hartwig and Dearling (1979) state we that we should be guided by principles of *scepticism* and *openness*; we ought to be sceptical to the possibility that any visualisation may obscure or misrepresent our data, and we should be open to the possibility of patterns and structures that we were not expecting

Some guiding principles for visualisation

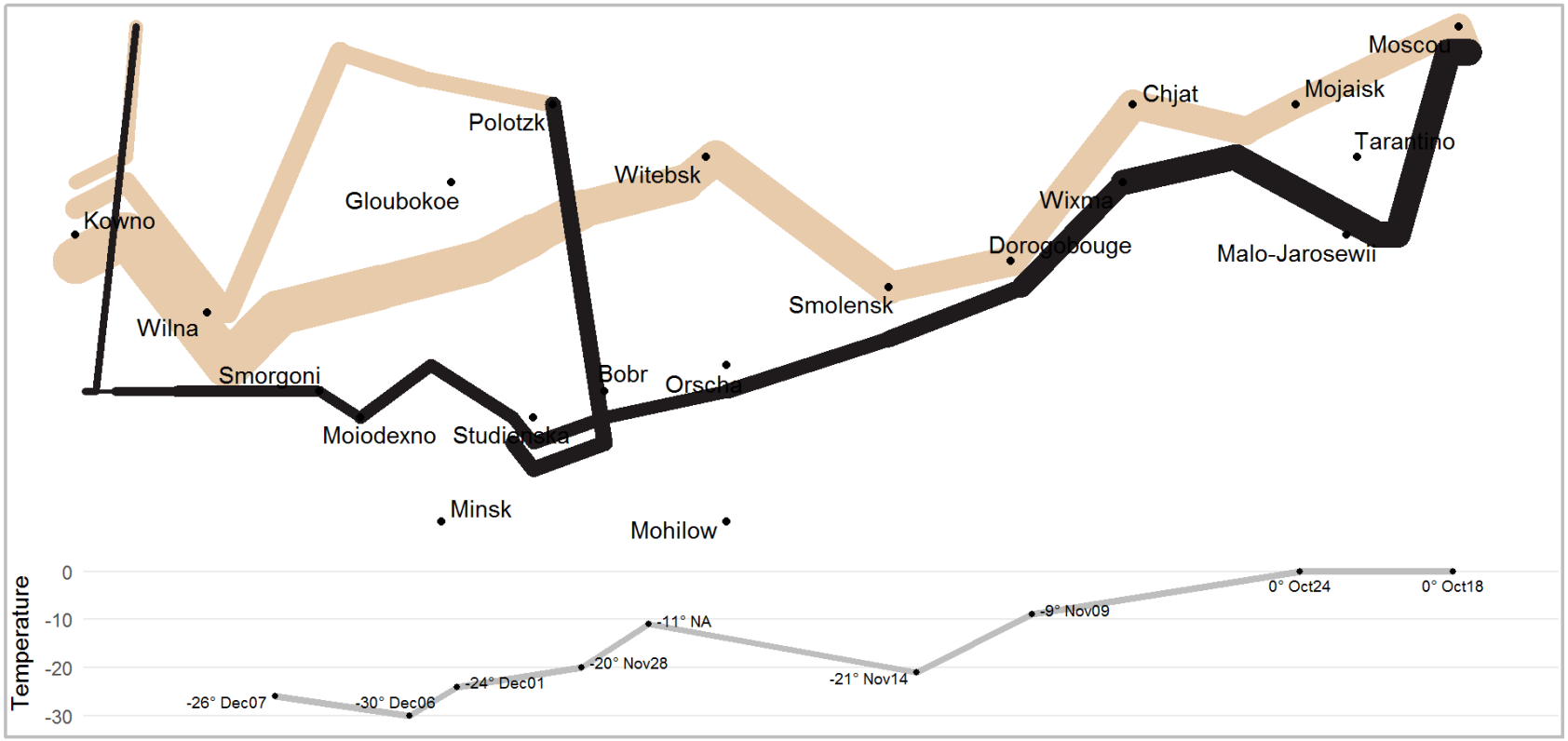
- Some guiding principles for visualisation mentioned by Edward R. Tufte in his *Visual Display of Quantitative Information* (Tufte, 1983) are the following:
  - *Above all else show the data*
  - *Avoid distorting what the data have to say*
  - *Present many numbers in a small space*
  - *Encourage the eye to compare different pieces of data*
  - *Reveal the data at several levels of detail, from a broad overview to the fine structure*

“The Best Statistical Graphic Ever Drawn”



Joseph Charles Minard, French civil engineer (1781 – 1870)

Recreated with ggplot



http://euclid.psych.yorku.ca/www/psy6135/tutorials/Minard.html

Some major visualisation tools

- *Histograms, density plots, bar plots*: These are used to display the distribution of values of continuous and discrete variables.
- *Boxplots*: Like histograms and density plots, boxplots (or box-and-whisker plots) display the distribution of values of continuous variables. However, they are more closely tied to robust statistical descriptions and so deserve to be treated as a class onto themselves.
- *Scatterplots*: Scatterplots and their variants such as *bubbleplots* are used to display bivariate data, or the relationships between two variables. Usually, scatterplots are used in cases where both variables are continuous, but may also be used, though perhaps with additional modification, when one variable is discrete.

Textbooks

- Atkinson, A. C. (1985) Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis.
- Healy, K. (2018) Data visualization: a practical introduction. Princeton University Press.
- Wickham, H. (2016) ggplot2: elegant graphics for data analysis. Springer.

Resources

- https://r-graph-gallery.com
- http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization