Data Visualization

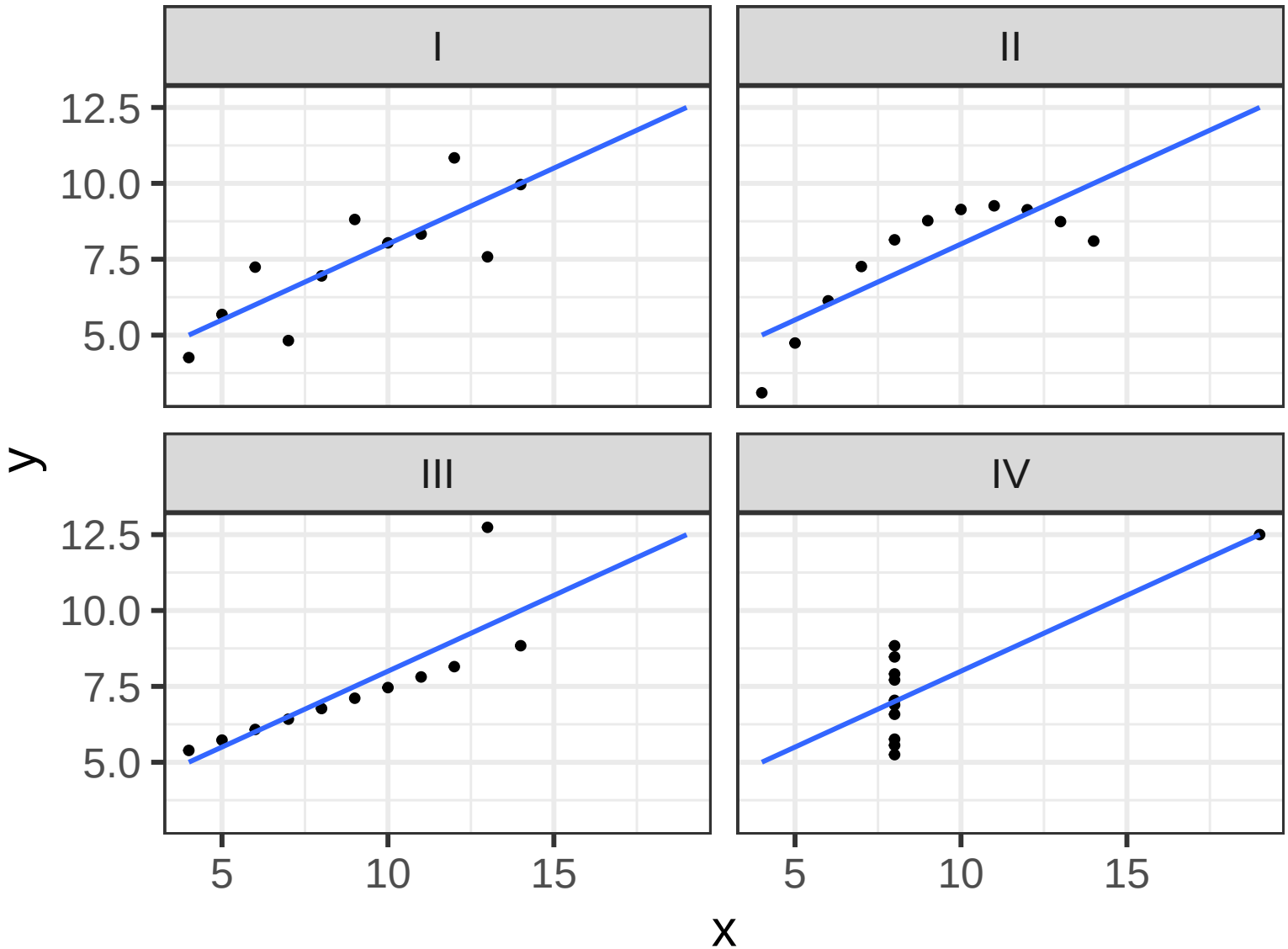
Mark Andrews
Psychology Department, Nottingham Trent University

mark.andrews@ntu.ac.uk

The aim of data visualization

- Rather than being a means to add some eye-candy or ornamentation to otherwise dull reports or slides, the prupose of visualization 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 *Anscombe's quartet* (Anscombe 1973):

set	mean(x)	mean(y)	sd(x)	sd(y)	cor(x, y)
I	9	7.5	3.32	2.03	0.82
II	9	7.5	3.32	2.03	0.82
III	9	7.5	3.32	2.03	0.82
IV	9	7.5	3.32	2.03	0.82


I
II
III
IV
12.5
10.0
7.5
5.0
5
10
15
y
x

The aim of data visualization

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

Some guiding principles for visualization 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*

Some major visualization 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.

References

Anscombe, Francis J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21.

Hartwig, Frederick, and Brian E Dearing. 1979. *Exploratory Data Analysis*. Sage.

Tufte, E. R. 1983. *The Visual Display of Quantitative Information*. Graphics Press.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Addison-Wesley Publishing Company.