Lancaster University

Quantitative Research Methods

Matthew Ivory

matthew.ivory@lancaster.ac.uk

1. Introduction to the course

- Session 1: Introduction to quantitative research methods using R
- Session 2: Data management and data wrangling
- Session 3: Exploratory data analysis
- **Session 4: Data visualization**
- Session 5: Live Coding Walkthrough
- Session 6: Probability and distributions
- Session 7: Tests for discrete variables: Analysing contingency tables
- Session 8: Correlations and t-tests
- Session 9: ANOVA and linear regression
- Session 10: Multiple regression, introduction to generalised linear regression

Our plan for today

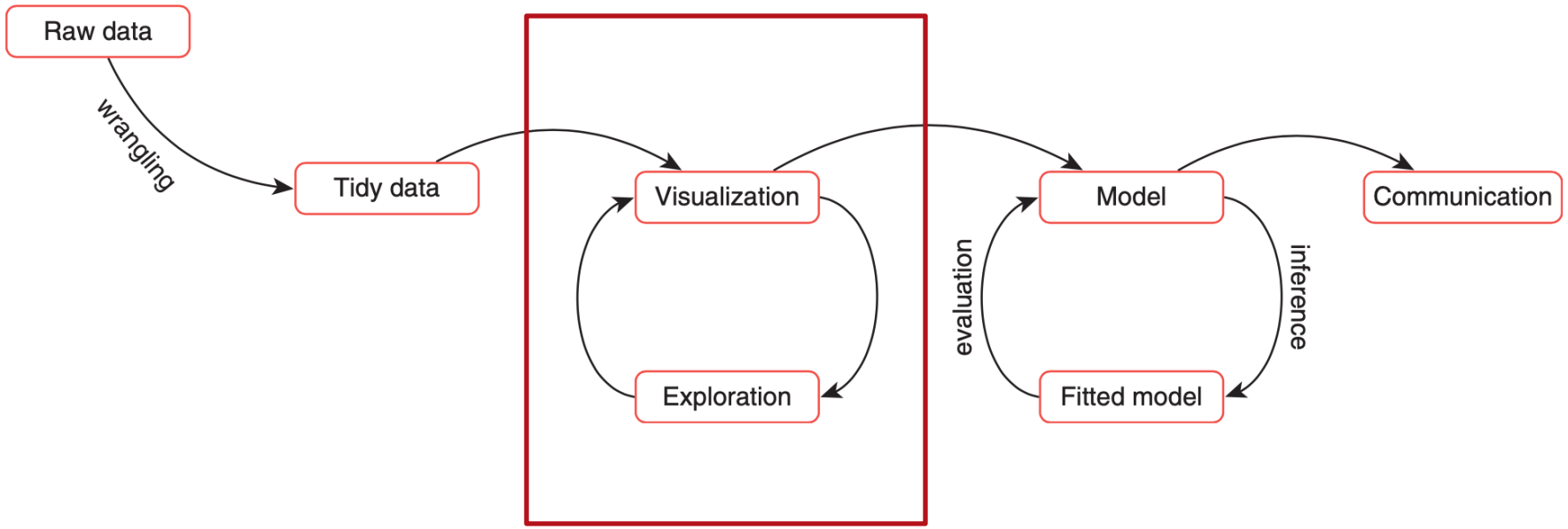
Exploratory data analysis (contd)

Data visualization

- Short introduction
- Handout 4

Exploratory data analysis

The Data Science Workflow



Reproduced from Andrews (2021)

Tukey (1977)

Exploratory data analysis

- Aim: to discover potentially interesting patterns and behaviors in the data.

Confirmatory data analysis:

- Aim: to propose and test models of our data

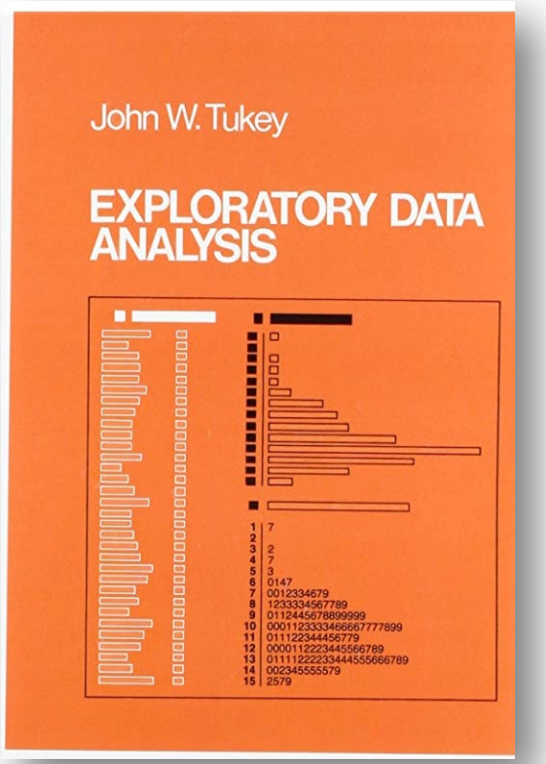
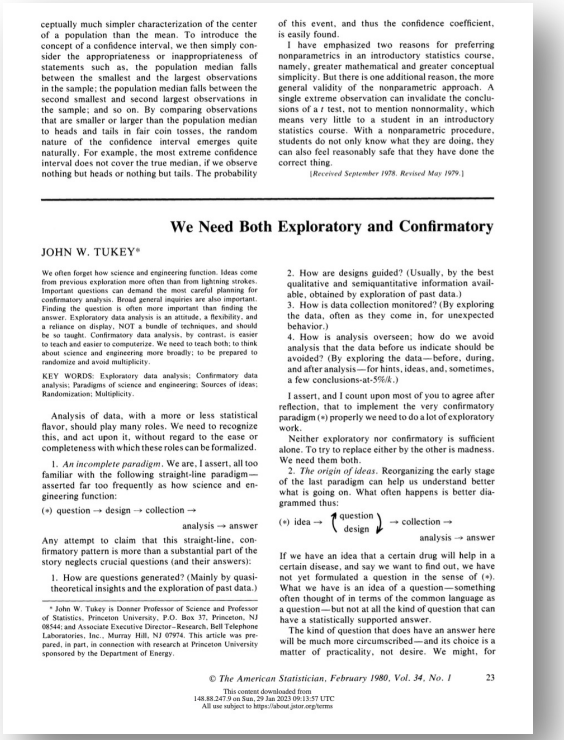


Image reproduced from here.

Tukey (1980)

"Exploratory data analysis is **an attitude, a flexibility, and a reliance on display**, NOT a bundle of techniques, and should be so taught.



ceptually much simpler characterization of the center of a population than the mean. To introduce the concept of a confidence interval, we then simply consider the appropriateness or inappropriateness of statements such as, the population median falls between the smallest and the largest observations in the sample; the population median falls between the second smallest and second largest observations in the sample; and so on. By comparing observations that are smaller or larger than the population median to heads and tails in fair coin tosses, the random nature of the confidence interval emerges quite naturally. For example, the most extreme confidence interval does not cover the true median, if we observe nothing but heads or nothing but tails. The probability of this event, and thus the confidence coefficient, is easily found.

I have emphasized two reasons for preferring nonparametrics in an introductory statistics course, namely, greater mathematical and greater conceptual simplicity. But there is one additional reason, the more general validity of the nonparametric approach. A single extreme observation can invalidate the conclusions of a t test, not to mention nonnormality, which means very little to a student in an introductory statistics course. With a nonparametric procedure, students do not only know what they are doing, they can also feel reasonably safe that they have done the correct thing.

[Received September 1978. Revised May 1979.]

We Need Both Exploratory and Confirmatory

JOHN W. TUKEY*

We often forget how science and engineering function. Ideas come from previous exploration more often than from lightning strokes. Important questions can demand the most careful planning for confirmatory analysis. Broad general inquiries are also important. Finding the question is often more important than finding the answer. Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should be so taught. Confirmatory data analysis, by contrast, is easier to teach and easier to computerize. We need to teach both; to think about science and engineering more broadly; to be prepared to randomize and avoid multiplicity.

KEY WORDS: Exploratory data analysis; Confirmatory data analysis; Paradigms of science and engineering; Sources of ideas; Randomization; Multiplicity.

Analysis of data, with a more or less statistical flavor, should play many roles. We need to recognize this, and act upon it, without regard to the ease or completeness with which these roles can be formalized.

1. *An incomplete paradigm.* We are, I assert, all too familiar with the following straight-line paradigm—asserted far too frequently as how science and engineering function:

(*) question → design → collection → analysis → answer

Any attempt to claim that this straight-line, confirmatory pattern is more than a substantial part of the story neglects crucial questions (and their answers):

1. How are questions generated? (Mainly by quasi-theoretical insights and the exploration of past data.)
2. How are designs guided? (Usually, by the best qualitative and semiquantitative information available, obtained by exploration of past data.)
3. How is data collection monitored? (By exploring the data, often as they come in, for unexpected behavior.)
4. How is analysis overseen; how do we avoid analysis that the data before us indicate should be avoided? (By exploring the data—before, during, and after analysis—for hints, ideas, and, sometimes, a few conclusions-at-5%/k.)

I assert, and I count upon most of you to agree after reflection, that to implement the very confirmatory paradigm (*) properly we need to do a lot of exploratory work.

Neither exploratory nor confirmatory is sufficient alone. To try to replace either by the other is madness. We need them both.

2. *The origin of ideas.* Reorganizing the early stage of the last paradigm can help us understand better what is going on. What often happens is better diagrammed thus:

(*) idea → { question / design } → collection → analysis → answer

If we have an idea that a certain drug will help in a certain disease, and say we want to find out, we have not yet formulated a question in the sense of (*). What we have is an idea of a question—something often thought of in terms of the common language as a question—but not at all the kind of question that can have a statistically supported answer.

The kind of question that does have an answer here will be much more circumscribed—and its choice is a matter of practicality, not desire. We might, for

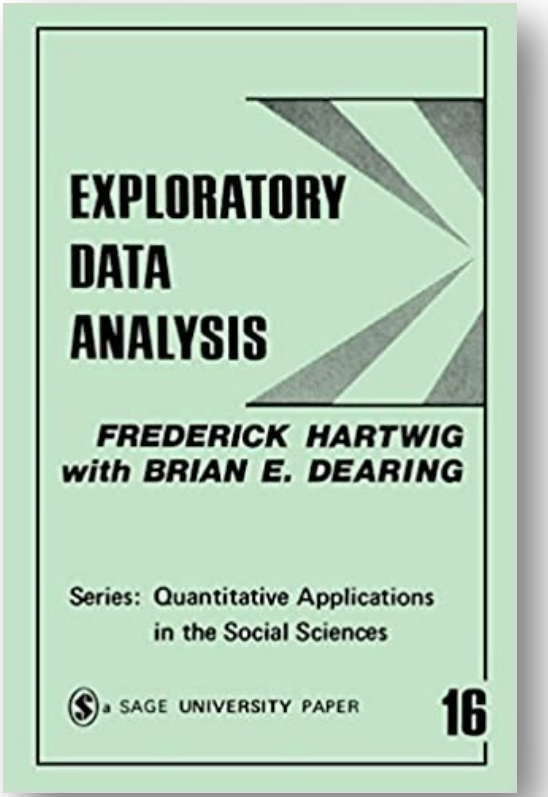
* John W. Tukey is Donner Professor of Science and Professor of Statistics, Princeton University, P.O. Box 37, Princeton, NJ 08544; and Associate Executive Director–Research, Bell Telephone Laboratories, Inc., Murray Hill, NJ 07974. This article was prepared, in part, in connection with research at Princeton University sponsored by the Department of Energy.

Hartwig & Dearling (1979)

"Exploratory data analysis is **a state of mind, a way of thinking about data analysis** (…)."

Assumption: The more we know about the data, the more effectively data can be used to develop, test and refine theory.

Examples of EDA in your field?



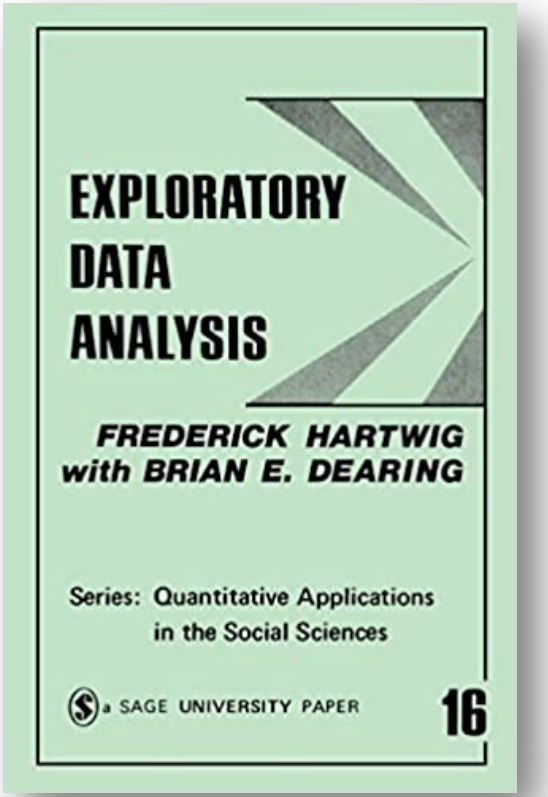
Hartwig & Dearling (1979)

The explanatory seeks to maximize what can be learned from data.

This requires two things:

1. **Skepticism**
2. **Openness**

What do you think this means?

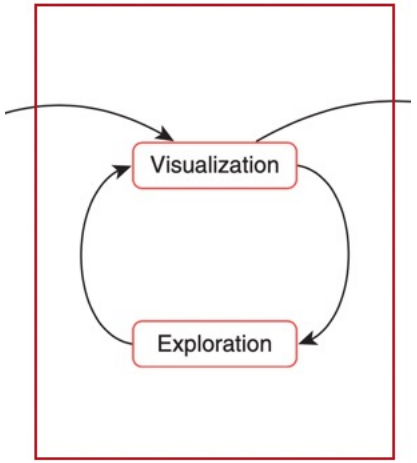


Data visualization

What is data visualization?

The graphic representation of data and information

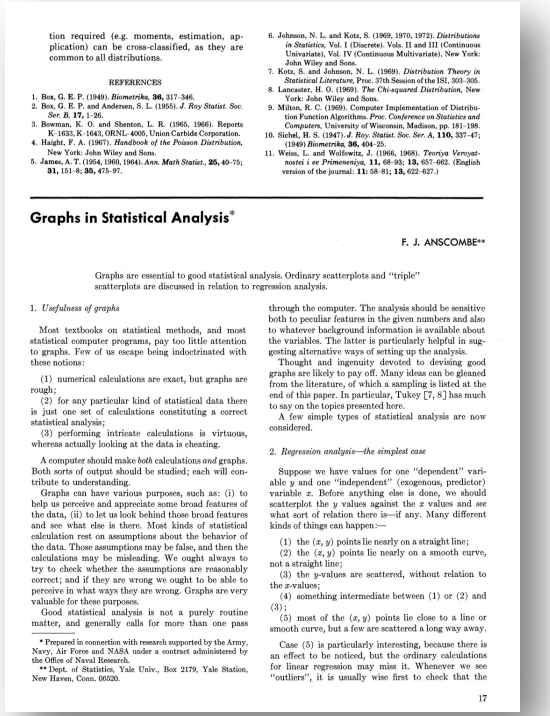
An essential component of EDA and CDA



Anscombe's Quartet

Anscombe (1973) illustrates beautifully why we cannot rely solely on numerical summaries.

Anscombe created **four data sets**, each with two variables (x, y).



tion required (e.g. moments, estimation, application) can be cross-classified, as they are common to all distributions.

REFERENCES

1. Box, G. E. P. (1949). *Biometrika*, **36**, 317–346.
2. Box, G. E. P. and Andersen, S. L. (1955). *J. Roy Statist. Soc. Ser. B*, **17**, 1–26.
3. Bowman, K. O. and Shenton, L. R. (1965, 1966). Reports K-1633, K-1643, ORNL-4005, Union Carbide Corporation.
4. Haight, F. A. (1967). *Handbook of the Poisson Distribution*, New York: John Wiley and Sons.
5. James, A. T. (1954, 1960, 1964). *Ann. Math Statist.*, **25**, 40–75; **31**, 151–8; **35**, 475–97.
6. Johnson, N. L. and Kotz, S. (1969, 1970, 1972). *Distributions in Statistics*, Vol. I (Discrete). Vols. II and III (Continuous Univariate), Vol. IV (Continuous Multivariate), New York: John Wiley and Sons.
7. Kotz, S. and Johnson, N. L. (1969). *Distribution Theory in Statistical Literature*, Proc. 37th Session of the ISI, 303–305.
8. Lancaster, H. O. (1969). *The Chi-squared Distribution*, New York: John Wiley and Sons.
9. Milton, R. C. (1969). Computer Implementation of Distribution Function Algorithms. *Proc. Conference on Statistics and Computers*, University of Wisconsin, Madison, pp. 181–198.
10. Sichel, H. S. (1947). *J. Roy. Statist. Soc. Ser. A*, **110**, 337–47; (1949) *Biometrika*, **36**, 404–25.
11. Weiss, L. and Wolfowitz, J. (1966, 1968). *Teoriya Veroyatnostei i ee Primeneniya*, **11**, 68–93; **13**, 657–662. (English version of the journal: **11**: 58–81; **13**, 622–627.)

Graphs in Statistical Analysis*

F. J. ANSCOMBE**

Graphs are essential to good statistical analysis. Ordinary scatterplots and "triple" scatterplots are discussed in relation to regression analysis.

1. *Usefulness of graphs*

Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:

(1) numerical calculations are exact, but graphs are rough;
(2) for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;
(3) performing intricate calculations is virtuous, whereas actually looking at the data is cheating.

A computer should make *both* calculations *and* graphs. Both sorts of output should be studied; each will contribute to understanding.

Graphs can have various purposes, such as: (i) to help us perceive and appreciate some broad features of the data, (ii) to let us look behind those broad features and see what else is there. Most kinds of statistical calculation rest on assumptions about the behavior of the data. Those assumptions may be false, and then the calculations may be misleading. We ought always to try to check whether the assumptions are reasonably correct; and if they are wrong we ought to be able to perceive in what ways they are wrong. Graphs are very valuable for these purposes.

Good statistical analysis is not a purely routine matter, and generally calls for more than one pass through the computer. The analysis should be sensitive both to peculiar features in the given numbers and also to whatever background information is available about the variables. The latter is particularly helpful in suggesting alternative ways of setting up the analysis.

Thought and ingenuity devoted to devising good graphs are likely to pay off. Many ideas can be gleaned from the literature, of which a sampling is listed at the end of this paper. In particular, Tukey [7, 8] has much to say on the topics presented here.

A few simple types of statistical analysis are now considered.

2. *Regression analysis—the simplest case*

Suppose we have values for one "dependent" variable y and one "independent" (exogenous, predictor) variable x. Before anything else is done, we should scatterplot the y values against the x values and see what sort of relation there is—if any. Many different kinds of things can happen:—

(1) the (x, y) points lie nearly on a straight line;
(2) the (x, y) points lie nearly on a smooth curve, not a straight line;
(3) the y-values are scattered, without relation to the x-values;
(4) something intermediate between (1) or (2) and (3);
(5) most of the (x, y) points lie close to a line or smooth curve, but a few are scattered a long way away.

Case (5) is particularly interesting, because there is an effect to be noticed, but the ordinary calculations for linear regression may miss it. Whenever we see "outliers", it is usually wise first to check that the

* Prepared in connection with research supported by the Army, Navy, Air Force and NASA under a contract administered by the Office of Naval Research.

** Dept. of Statistics, Yale Univ., Box 2179, Yale Station, New Haven, Conn. 06520.

17

Data set		1-3	1	2	3		4	4
Variable		x	y	y	y		x	y
Obs. no.	1	10.0	8.04	9.14	7.46		8.0	6.58
	2	8.0	6.95	8.14	6.77		8.0	5.76
	3	13.0	7.58	8.74	12.74		8.0	7.71
	4	9.0	8.81	8.77	7.11		8.0	8.84
	5	11.0	8.33	9.26	7.81		8.0	8.47
	6	14.0	9.96	8.10	8.84		8.0	7.04
	7	6.0	7.24	6.13	6.08		8.0	5.25
	8	4.0	4.26	3.10	5.39		19.0	12.50
	9	12.0	10.84	9.13	8.15		8.0	5.56
	10	7.0	4.82	7.26	6.42		8.0	7.91
	11	5.0	5.68	4.74	5.73		8.0	6.89

TABLE. Four data sets, each comprising 11 (x, y) pairs.

Data sets

The summary statistics for the data sets above

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

The numerical summary statistics would suggest that the data in these sets would be very similar!

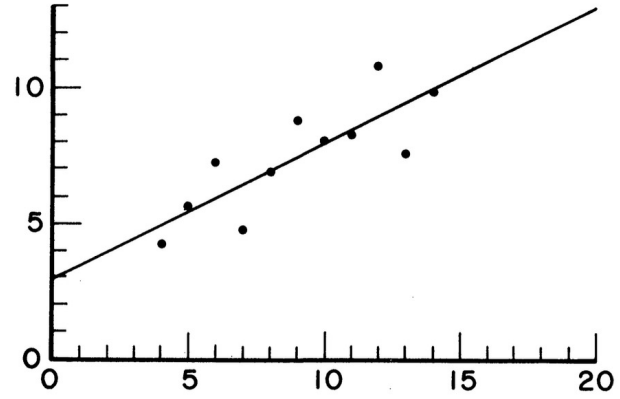


Figure 1

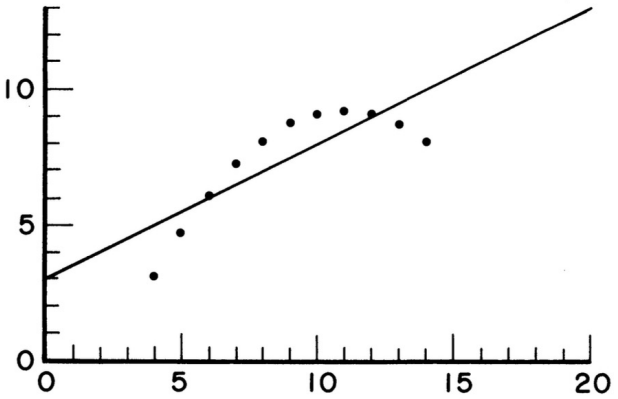


Figure 2

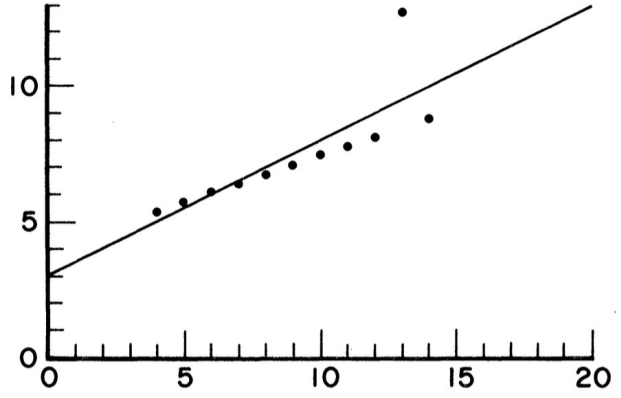


Figure 3

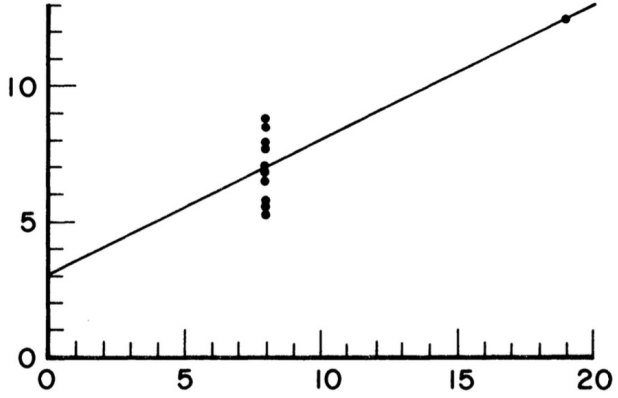
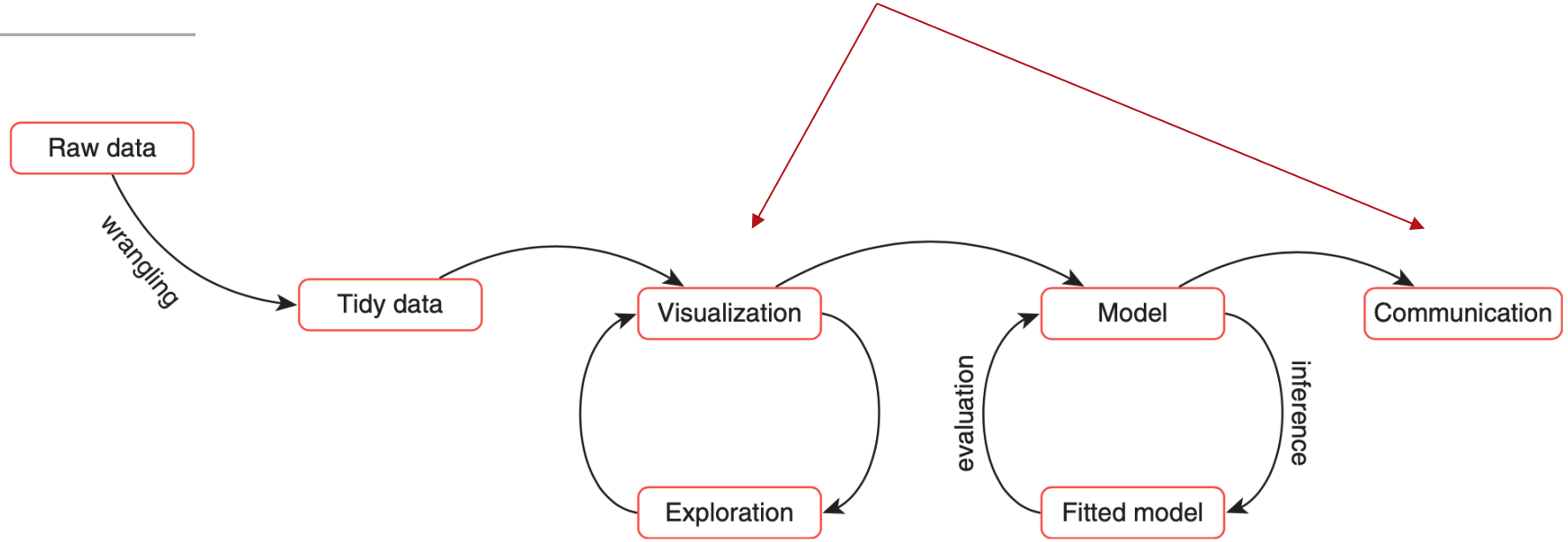


Figure 4

Tukey (1977): Visualization “forces us to notice what we never expected to see.”

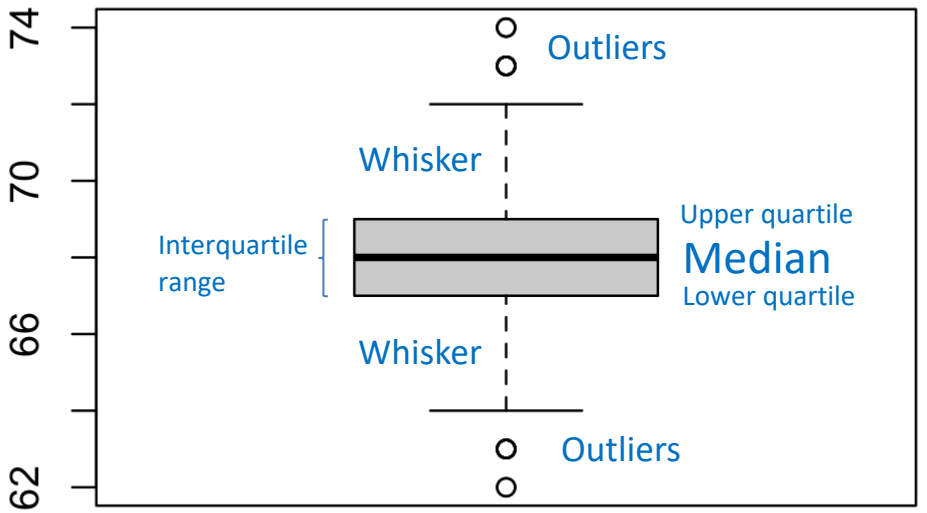
We use visualization twice
(at least)



Reproduced from Andrews (2021)

Data visualization in EDA

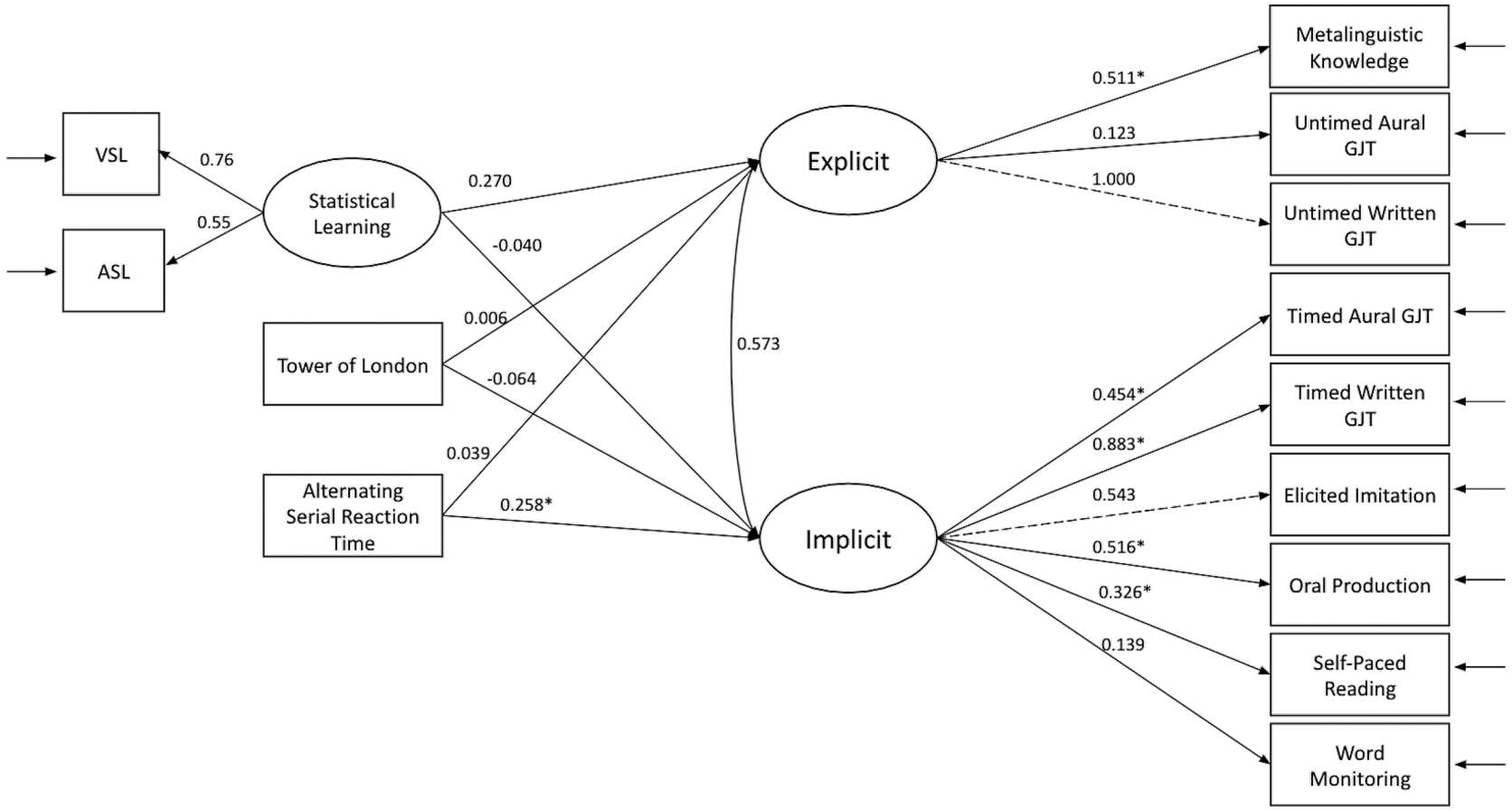
- Here, visualization allows us to explore data and find patterns that we can easily miss if we just rely on numerical summary statistics.



Data visualization in CDA

- We use graphics in publications and presentations.
- There are numeric techniques which are almost always displayed in visual form. Usually applies to techniques that model multivariate relationships.
- Example: The results of structural equation modeling (SME) are usually represented in the form of dendrograms.

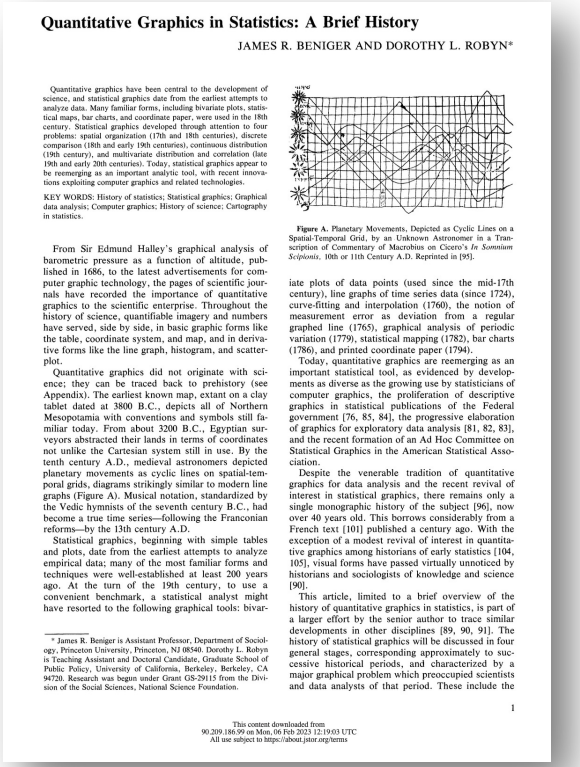
Godfroid and Kim (2021)



Historical perspective

Statistical graphics have been around for many centuries.

- c. 3,800 BC: Oldest known map (Northern Mesopotamia) on clay tablet
- 1786: Bar chart (William Playfair)
- 1801: Pie chart (William Playfair)
- 1833: Histogram (A. M. Guerry)
- 1874: Age pyramid (bilateral histogram) (Francis A. Walker)



Quantitative Graphics in Statistics: A Brief History

JAMES R. BENIGER AND DOROTHY L. ROBYN*

Quantitative graphics have been central to the development of science, and statistical graphics date from the earliest attempts to analyze data. Many familiar forms, including bivariate plots, statistical maps, bar charts, and coordinate paper, were used in the 18th century. Statistical graphics developed through attention to four problems: spatial organization (17th and 18th centuries), discrete comparison (18th and early 19th centuries), continuous distribution (19th century), and multivariate distribution and correlation (late 19th and early 20th centuries). Today, statistical graphics appear to be reemerging as an important analytic tool, with recent innovations exploiting computer graphics and related technologies.

KEY WORDS: History of statistics; Statistical graphics; Graphical data analysis; Computer graphics; History of science; Cartography in statistics.

From Sir Edmund Halley's graphical analysis of barometric pressure as a function of altitude, published in 1686, to the latest advertisements for computer graphic technology, the pages of scientific journals have recorded the importance of quantitative graphics to the scientific enterprise. Throughout the history of science, quantifiable imagery and numbers have served, side by side, in basic graphic forms like the table, coordinate system, and map, and in derivative forms like the line graph, histogram, and scatterplot.

Quantitative graphics did not originate with science; they can be traced back to prehistory (see Appendix). The earliest known map, extant on a clay tablet dated at 3800 B.C., depicts all of Northern Mesopotamia with conventions and symbols still familiar today. From about 3200 B.C., Egyptian surveyors abstracted their lands in terms of coordinates not unlike the Cartesian system still in use. By the tenth century A.D., medieval astronomers depicted planetary movements as cyclic lines on spatial-temporal grids, diagrams strikingly similar to modern line graphs (Figure A). Musical notation, standardized by the Vedic hymnists of the seventh century B.C., had become a true time series—following the Franconian reforms—by the 13th century A.D.

Statistical graphics, beginning with simple tables and plots, date from the earliest attempts to analyze empirical data; many of the most familiar forms and techniques were well-established at least 200 years ago. At the turn of the 19th century, to use a convenient benchmark, a statistical analyst might have resorted to the following graphical tools: bivariate plots of data points (used since the mid-17th century), line graphs of time series data (since 1724), curve-fitting and interpolation (1760), the notion of measurement error as deviation from a regular graphed line (1765), graphical analysis of periodic variation (1779), statistical mapping (1782), bar charts (1786), and printed coordinate paper (1794).

Today, quantitative graphics are reemerging as an important statistical tool, as evidenced by developments as diverse as the growing use by statisticians of computer graphics, the proliferation of descriptive graphics in statistical publications of the Federal government [76, 85, 84], the progressive elaboration of graphics for exploratory data analysis [81, 82, 83], and the recent formation of an Ad Hoc Committee on Statistical Graphics in the American Statistical Association.

Despite the venerable tradition of quantitative graphics for data analysis and the recent revival of interest in statistical graphics, there remains only a single monographic history of the subject [96], now over 40 years old. This borrows considerably from a French text [101] published a century ago. With the exception of a modest revival of interest in quantitative graphics among historians of early statistics [104, 105], visual forms have passed virtually unnoticed by historians and sociologists of knowledge and science [90].

This article, limited to a brief overview of the history of quantitative graphics in statistics, is part of a larger effort by the senior author to trace similar developments in other disciplines [89, 90, 91]. The history of statistical graphics will be discussed in four general stages, corresponding approximately to successive historical periods, and characterized by a major graphical problem which preoccupied scientists and data analysts of that period. These include the

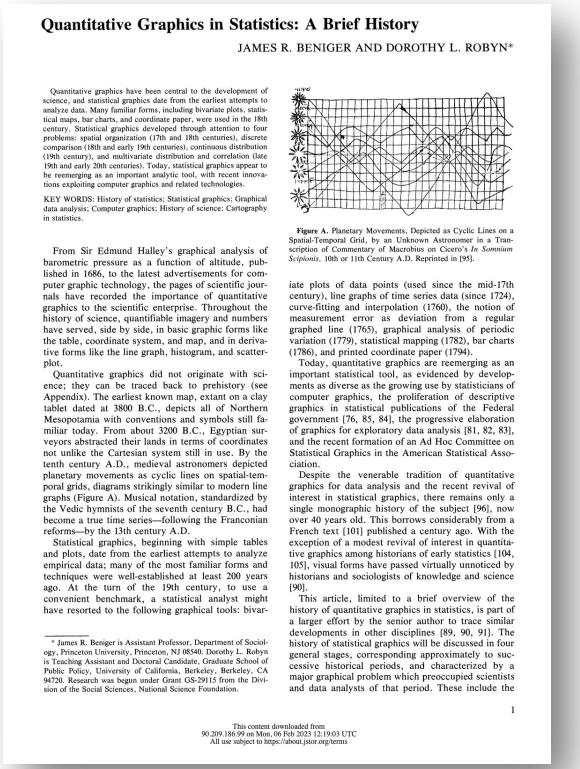
Figure A. Planetary Movements, Depicted as Cyclic Lines on a Spatial-Temporal Grid, by an Unknown Astronomer in a Transcription of Commentary of Macrobius on Cicero's *In Somnium Scipionis*, 10th or 11th Century A.D. Reprinted in [95].

* James R. Beniger is Assistant Professor, Department of Sociology, Princeton University, Princeton, NJ 08540. Dorothy L. Robyn is Teaching Assistant and Doctoral Candidate, Graduate School of Public Policy, University of California, Berkeley, Berkeley, CA 94720. Research was begun under Grant GS-29115 from the Division of the Social Sciences, National Science Foundation.

1

Beninger and Robyn (1978)

Development of statistical graphics as solutions to major graphical problems that occupied scientists at the time.



Quantitative Graphics in Statistics: A Brief History

JAMES R. BENIGER AND DOROTHY L. ROBYN*

Quantitative graphics have been central to the development of science, and statistical graphics date from the earliest attempts to analyze data. Many familiar forms, including bivariate plots, statistical maps, bar charts, and coordinate paper, were used in the 18th century. Statistical graphics developed through attention to four problems: spatial organization (17th and 18th centuries), discrete comparison (18th and early 19th centuries), continuous distribution (19th century), and multivariate distribution and correlation (late 19th and early 20th centuries). Today, statistical graphics appear to be reemerging as an important analytic tool, with recent innovations exploiting computer graphics and related technologies.

KEY WORDS: History of statistics; Statistical graphics; Graphical data analysis; Computer graphics; History of science; Cartography in statistics.

From Sir Edmund Halley's graphical analysis of barometric pressure as a function of altitude, published in 1686, to the latest advertisements for computer graphic technology, the pages of scientific journals have recorded the importance of quantitative graphics to the scientific enterprise. Throughout the history of science, quantifiable imagery and numbers have served, side by side, in basic graphic forms like the table, coordinate system, and map, and in derivative forms like the line graph, histogram, and scatterplot.

Quantitative graphics did not originate with science; they can be traced back to prehistory (see Appendix). The earliest known map, extant on a clay tablet dated at 3800 B.C., depicts all of Northern Mesopotamia with conventions and symbols still familiar today. From about 3200 B.C., Egyptian surveyors abstracted their lands in terms of coordinates not unlike the Cartesian system still in use. By the tenth century A.D., medieval astronomers depicted planetary movements as cyclic lines on spatial-temporal grids, diagrams strikingly similar to modern line graphs (Figure A). Musical notation, standardized by the Vedic hymnists of the seventh century B.C., had become a true time series—following the Franconian reforms—by the 13th century A.D.

Statistical graphics, beginning with simple tables and plots, date from the earliest attempts to analyze empirical data; many of the most familiar forms and techniques were well-established at least 200 years ago. At the turn of the 19th century, to use a convenient benchmark, a statistical analyst might have resorted to the following graphical tools: bivariate plots of data points (used since the mid-17th century), line graphs of time series data (since 1724), curve-fitting and interpolation (1760), the notion of measurement error as deviation from a regular graphed line (1765), graphical analysis of periodic variation (1779), statistical mapping (1782), bar charts (1786), and printed coordinate paper (1794).

Today, quantitative graphics are reemerging as an important statistical tool, as evidenced by developments as diverse as the growing use by statisticians of computer graphics, the proliferation of descriptive graphics in statistical publications of the Federal government [76, 85, 84], the progressive elaboration of graphics for exploratory data analysis [81, 82, 83], and the recent formation of an Ad Hoc Committee on Statistical Graphics in the American Statistical Association.

Despite the venerable tradition of quantitative graphics for data analysis and the recent revival of interest in statistical graphics, there remains only a single monographic history of the subject [96], now over 40 years old. This borrows considerably from a French text [101] published a century ago. With the exception of a modest revival of interest in quantitative graphics among historians of early statistics [104, 105], visual forms have passed virtually unnoticed by historians and sociologists of knowledge and science [90].

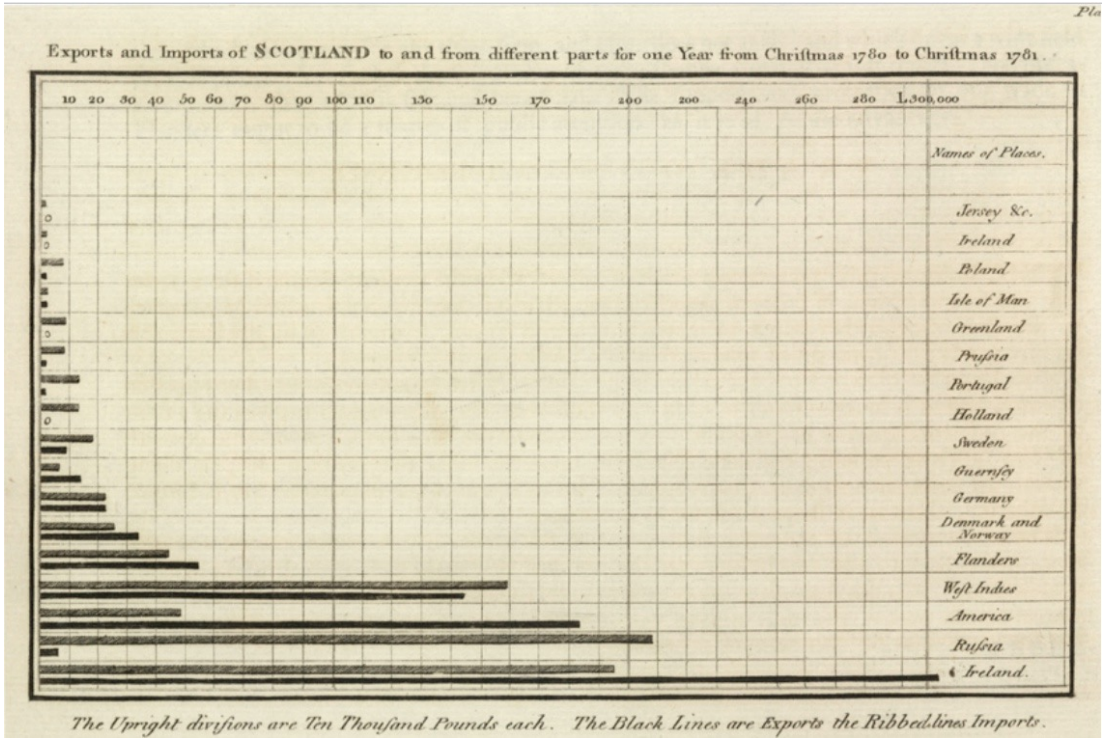
This article, limited to a brief overview of the history of quantitative graphics in statistics, is part of a larger effort by the senior author to trace similar developments in other disciplines [89, 90, 91]. The history of statistical graphics will be discussed in four general stages, corresponding approximately to successive historical periods, and characterized by a major graphical problem which preoccupied scientists and data analysts of that period. These include the

Figure A. Planetary Movements, Depicted as Cyclic Lines on a Spatial-Temporal Grid, by an Unknown Astronomer in a Transcription of Commentary of Macrobius on Cicero's *In Somnium Scipionis*, 10th or 11th Century A.D. Reprinted in [95].

* James R. Beniger is Assistant Professor, Department of Sociology, Princeton University, Princeton, NJ 08540. Dorothy L. Robyn is Teaching Assistant and Doctoral Candidate, Graduate School of Public Policy, University of California, Berkeley, Berkeley, CA 94720. Research was begun under Grant GS-29115 from the Division of the Social Sciences, National Science Foundation.

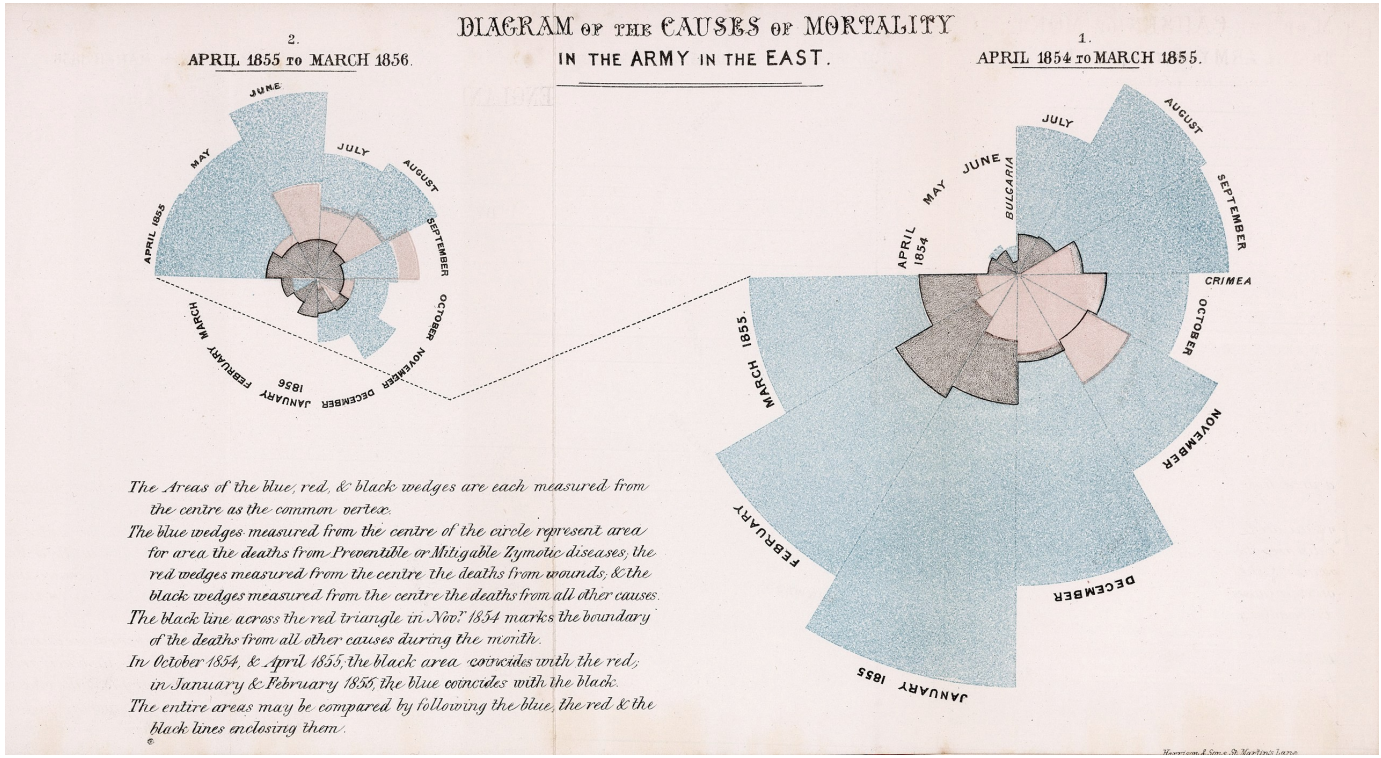
1

18th century: How to represent State statistics?



William Playfair (1786): Exports and Imports of Scotland to and from different parts for one Year from Christmas 1780 to Christmas 1781

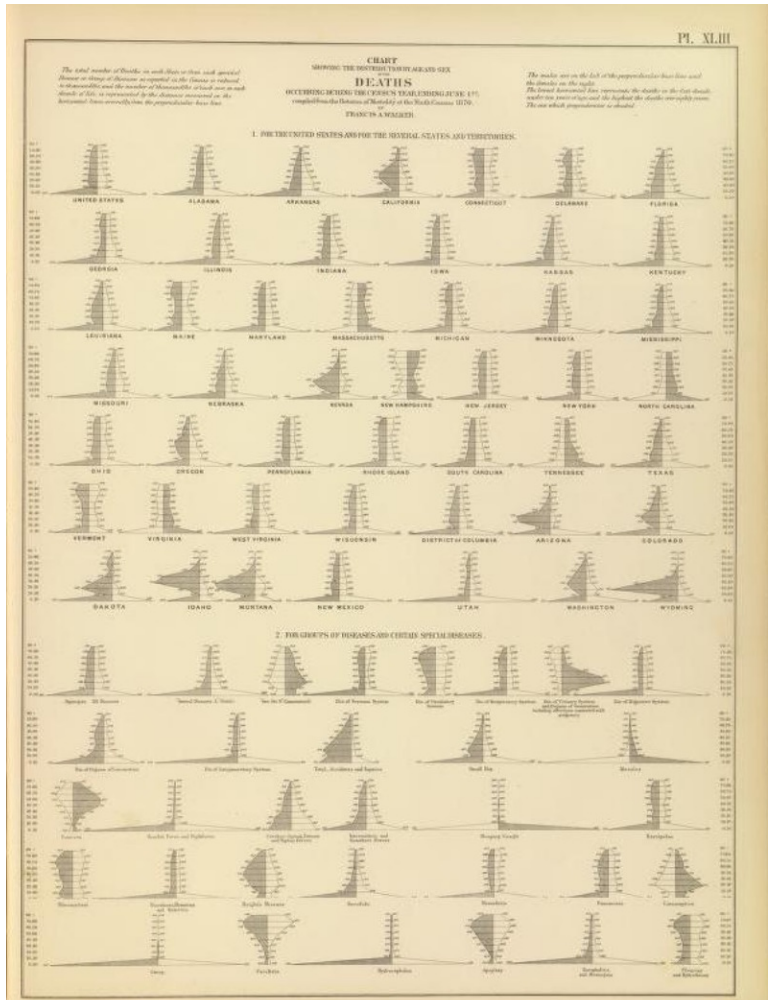
19th century: How to represent vital statistics?



Florence Nightingale (1858): Diagram of the causes of mortality in the army in the East

19th century: How to represent vital statistics?

Francis A. Walker (1874):
Age pyramid
(bilateral histogram)



Choose an area

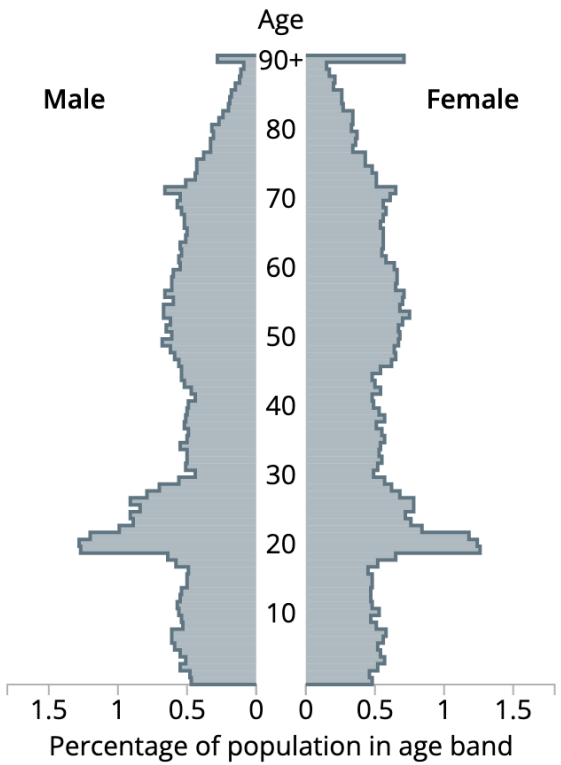
Lancaster

144,246 people in 2018

All ages

71,233 males 49.4%

73,013 females 50.6%



Choose an area

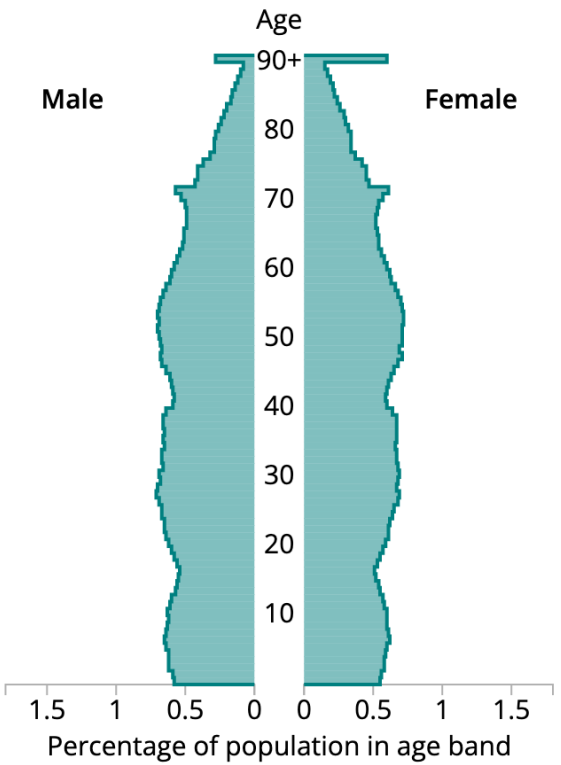
United Kingdom

66,435,550 people in 2018

All ages

32,790,202 males 49.4%

33,645,348 females 50.6%



ONS data, click here

19-20th century: How to represent multivariate relationships?

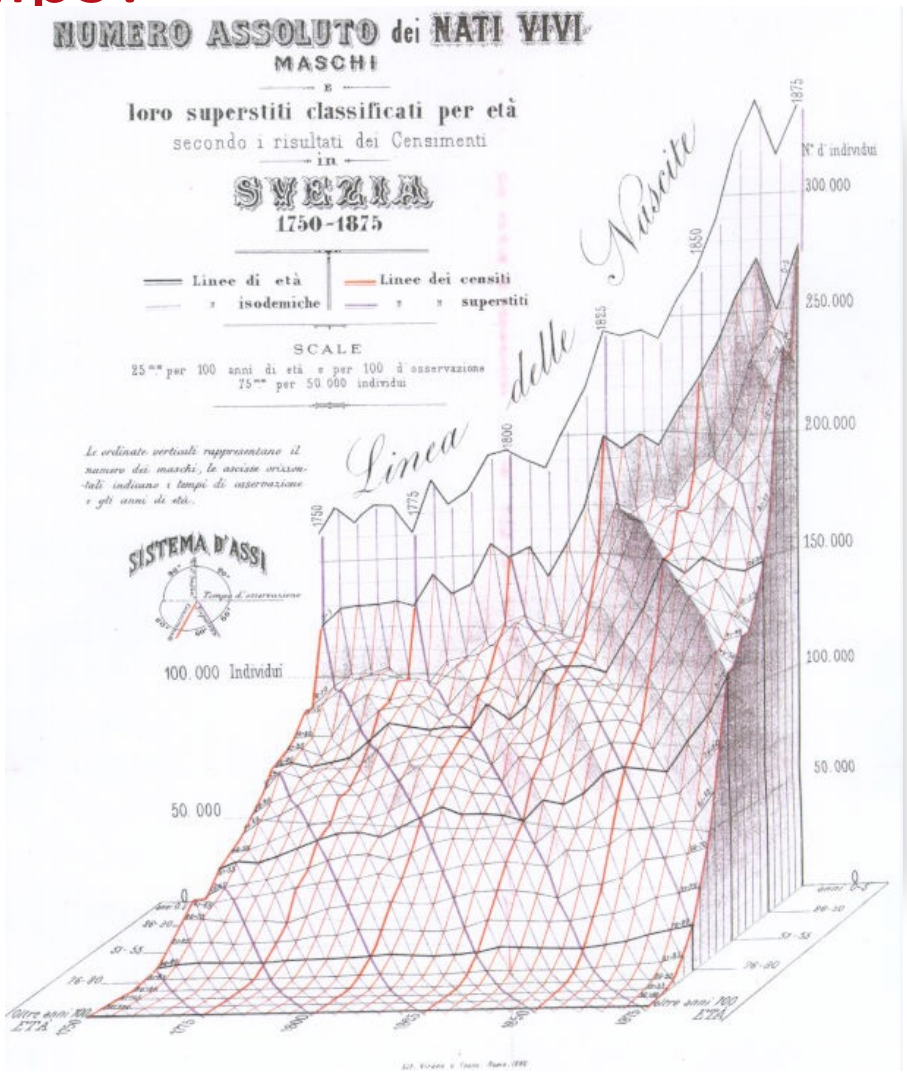
Vauthier (1874): Population contour map.

Map of Paris with population density depicted as contour lines.

19-20th century: How to represent multivariate relationships?

Luigi Perozzo (1879): Stereogram

3D population pyramid representing Swedish population (1750-1875) by age groups)

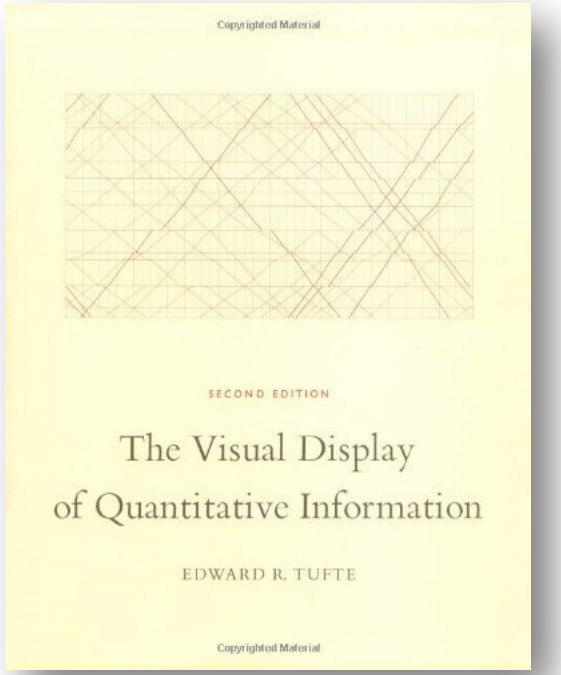


Some guiding principles

Look at your data from different ways and from different perspectives, compare finer and coarser levels.

Tufte (1983): Visual display of quantitative information

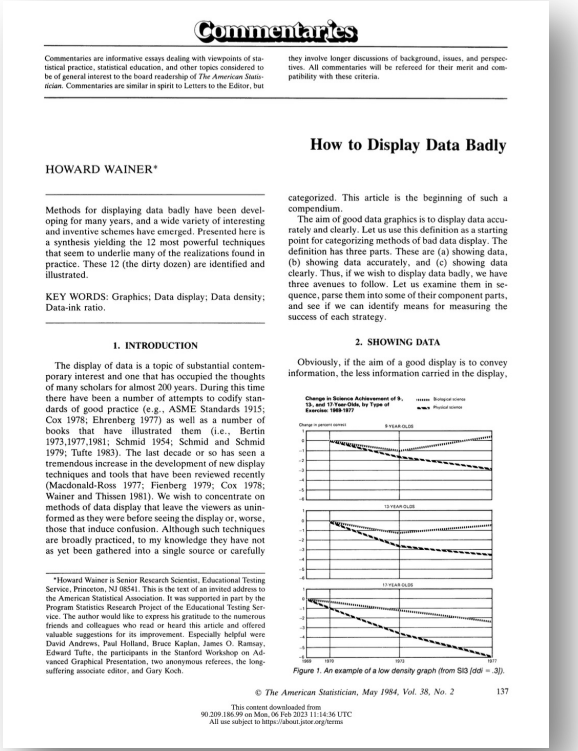
- Above all else show the data
- Avoid distorting what the data have to say
- Encourage the eye to compare different pieces of data
- Reveal the data at several levels of detail, from a broad overview to the fine structure



Wainer (1984): How to display data badly

Among the 12 rules:

- Show as few data as possible (minimize the data density)
- Hide what data you want to show
- Change scales in mid-axis
- Emphasize the trivial (ignore the important)
- If it has been done well in the past, think of another way to do it



Commentaries

Commentaries are informative essays dealing with viewpoints of statistical practice, statistical education, and other topics considered to be of general interest to the board readership of *The American Statistician*. Commentaries are similar in spirit to Letters to the Editor, but they involve longer discussions of background, issues, and perspectives. All commentaries will be refereed for their merit and compatibility with these criteria.

How to Display Data Badly

HOWARD WAINER*

Methods for displaying data badly have been developing for many years, and a wide variety of interesting and inventive schemes have emerged. Presented here is a synthesis yielding the 12 most powerful techniques that seem to underlie many of the realizations found in practice. These 12 (the dirty dozen) are identified and illustrated.

KEY WORDS: Graphics; Data display; Data density; Data-ink ratio.

1. INTRODUCTION

The display of data is a topic of substantial contemporary interest and one that has occupied the thoughts of many scholars for almost 200 years. During this time there have been a number of attempts to codify standards of good practice (e.g., ASME Standards 1915; Cox 1978; Ehrenberg 1977) as well as a number of books that have illustrated them (i.e., Bertin 1973,1977,1981; Schmid 1954; Schmid and Schmid 1979; Tufte 1983). The last decade or so has seen a tremendous increase in the development of new display techniques and tools that have been reviewed recently (Macdonald-Ross 1977; Fienberg 1979; Cox 1978; Wainer and Thissen 1981). We wish to concentrate on methods of data display that leave the viewers as uninformed as they were before seeing the display or, worse, those that induce confusion. Although such techniques are broadly practiced, to my knowledge they have not as yet been gathered into a single source or carefully categorized. This article is the beginning of such a compendium.

The aim of good data graphics is to display data accurately and clearly. Let us use this definition as a starting point for categorizing methods of bad data display. The definition has three parts. These are (a) showing data, (b) showing data accurately, and (c) showing data clearly. Thus, if we wish to display data badly, we have three avenues to follow. Let us examine them in sequence, parse them into some of their component parts, and see if we can identify means for measuring the success of each strategy.

2. SHOWING DATA

Obviously, if the aim of a good display is to convey information, the less information carried in the display,

*Howard Wainer is Senior Research Scientist, Educational Testing Service, Princeton, NJ 08541. This is the text of an invited address to the American Statistical Association. It was supported in part by the Program Statistics Research Project of the Educational Testing Service. The author would like to express his gratitude to the numerous friends and colleagues who read or heard this article and offered valuable suggestions for its improvement. Especially helpful were David Andrews, Paul Holland, Bruce Kaplan, James O. Ramsay, Edward Tufte, the participants in the Stanford Workshop on Advanced Graphical Presentation, two anonymous referees, the long-suffering associate editor, and Gary Koch.

Figure 1. An example of a low density graph (from SIS [ddi = .3]).

Wainer (1984)

- Data density index (ddi): The number of numbers plotted per square inch." (Tufte, 1983)

Show as few data as possible (minimize the data density)

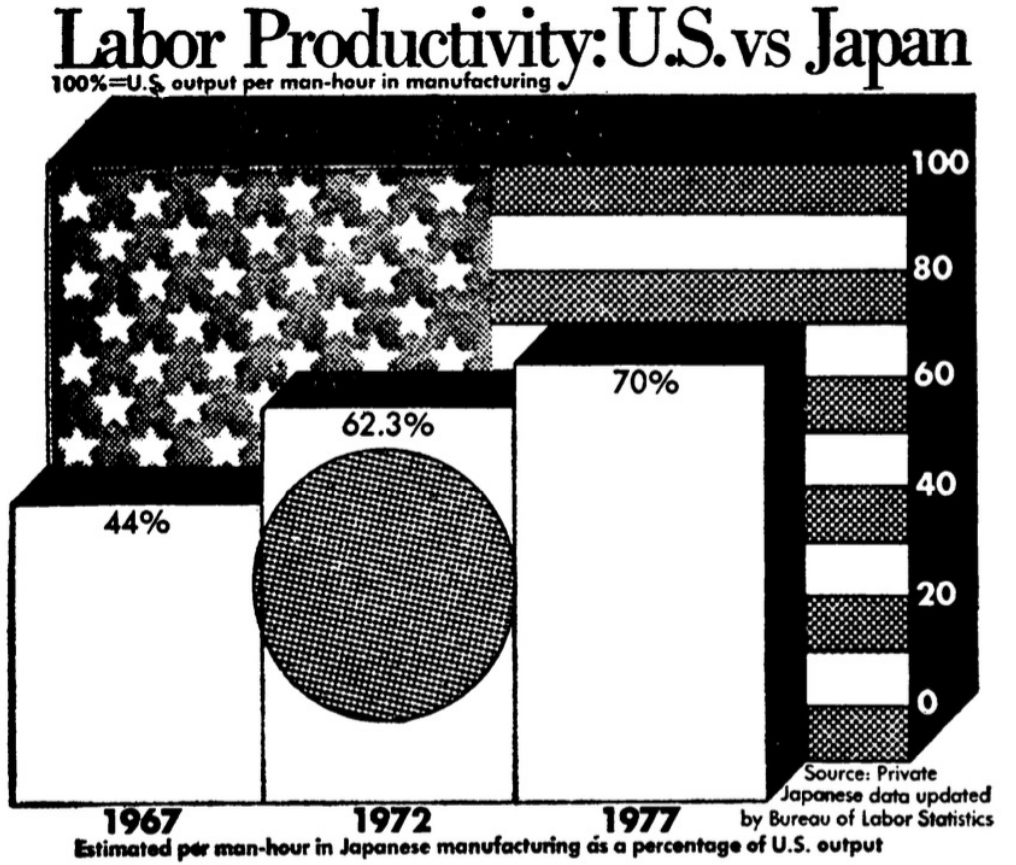


Figure 3. A low density graph (© 1978, The Washington Post) with chart-junk to fill in the space (ddi = .2).

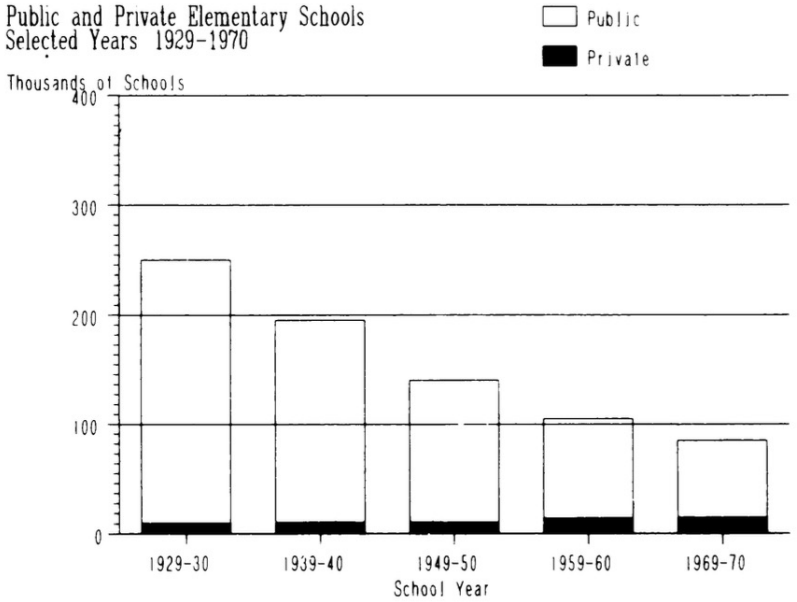


Figure 4. Hiding the data in the scale (from SI3).

Hide what data you want to show

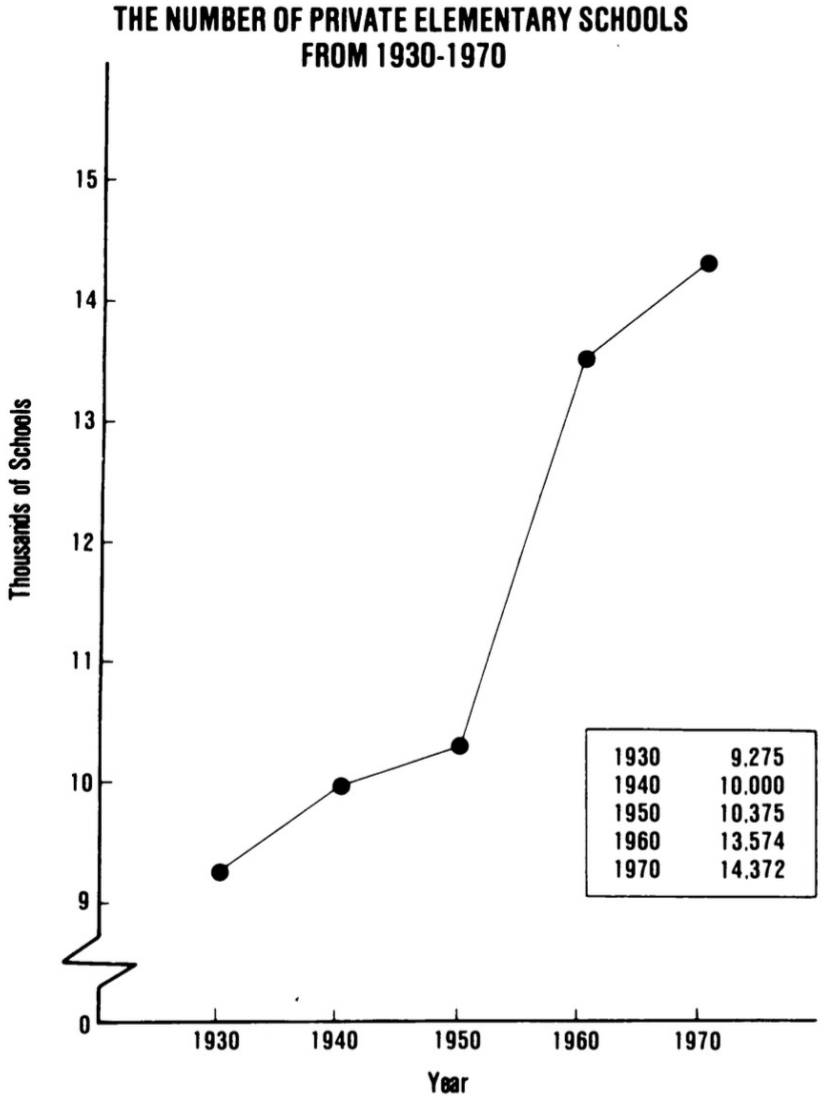


Figure 5. Expanding the scale and showing the data in Figure 4 (from SI3).

Wainer (1984)

Change scales in mid-axis

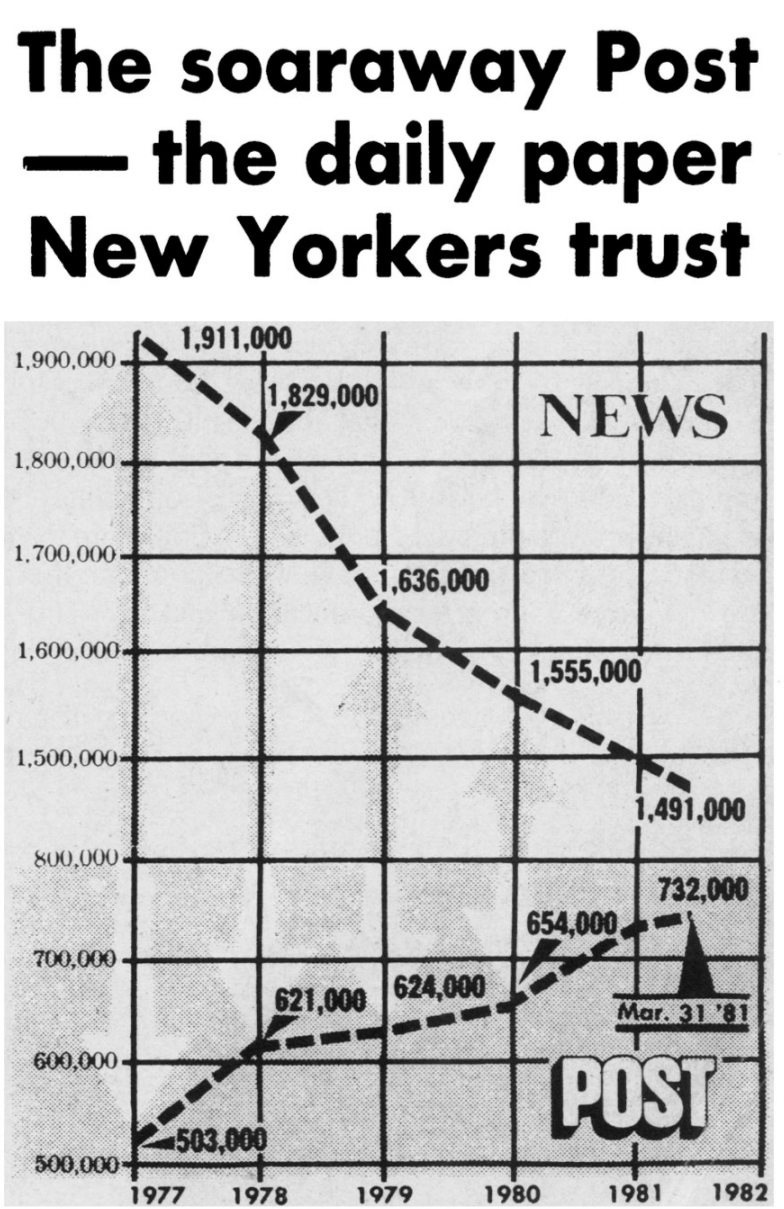


Figure 12. Changing scale in mid-axis to make large differences small (© *1981,* New York Post).

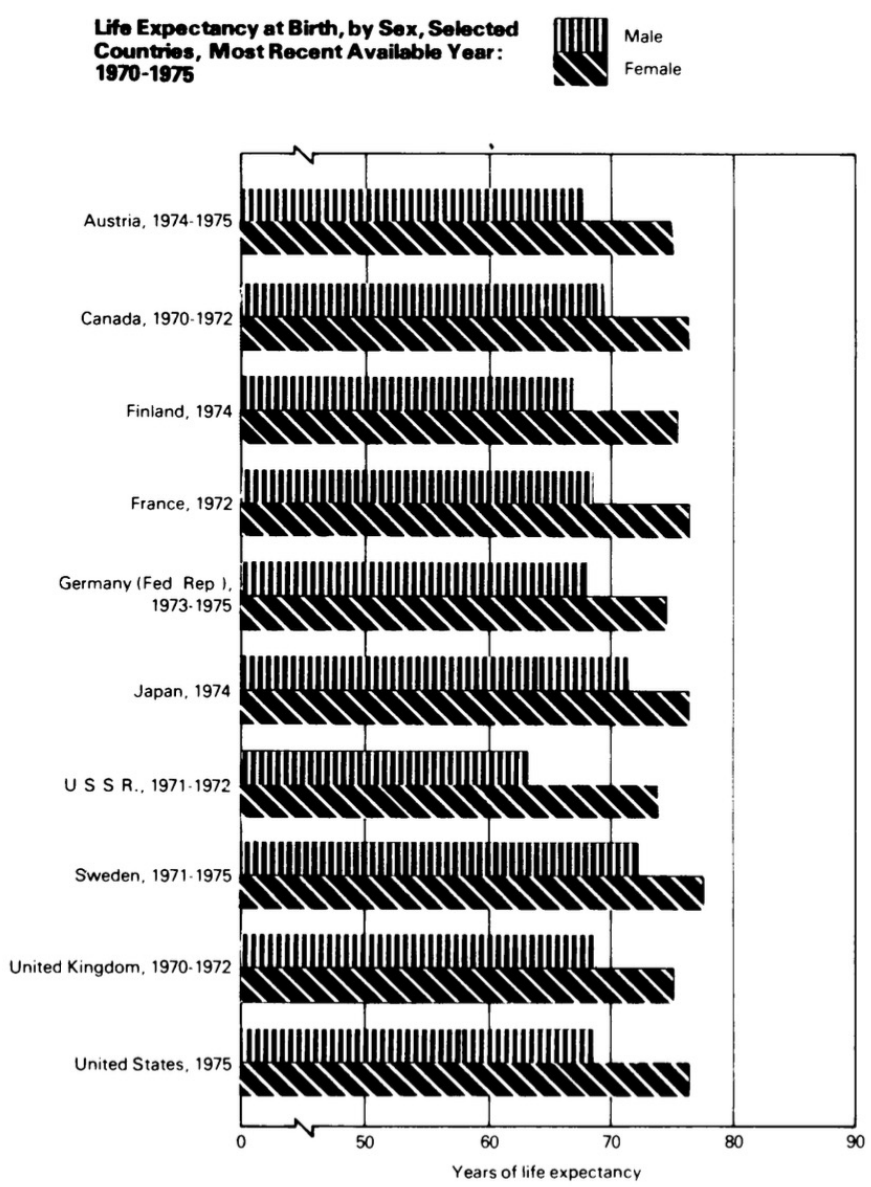


Figure 19. Austria First! Obscuring the data structure by alphabetizing the plot (from SI3).

LIFE EXPECTANCY AT BIRTH, BY SEX,
MOST RECENT AVAILABLE YEAR

WOMEN	YEARS	MEN
SWEDEN	78	
	77	
FRANCE, US, JAPAN, CANADA	76	
FINLAND, AUSTRIA, UK	75	
USSR, GERMANY	74	
	73	
	72	SWEDEN
	71	JAPAN
	70	
	69	CANADA, UK, US, FRANCE
	68	GERMANY, AUSTRIA
	67	FINLAND
	66	
	65	
	64	
	63	USSR
	62	

Figure 20. Ordering and spacing the data from Figure 19 as a stem-and-leaf diagram provides insights previously difficult to extract (from SI3).

Lancaster University

Some common type of graphics

Some common type of graphics

The R Graph Gallery

Welcome the R graph gallery, a collection of charts made with the R programming language. Hundreds of charts are displayed in several sections, always with their reproducible code available. The gallery makes a focus on the tidyverse and ggplot2. Feel free to suggest a chart or report a bug; any feedback is highly welcome. Stay in touch with the gallery by following it on Twitter or Github. If you're new to R, consider following this course.

Some common types of graphics

- Bar chart (simple and stacked)
- Pie charts
- Histograms
- Scatterplots
- Line graphs
- Box plots

A few practicalities

Cost and use of color

- Some publishers charge you $$$$ for including graphics, so read the “small print” carefully before you submitting your manuscript.
- Color looks nice on screen, but it doesn’t photocopy or print well.
- Also, publishers also often charge extra $$$$ for colored graphics.
- So, it is often best to stick to monochrome fill patterns in these cases.

Clear labeling

Always label your graphs carefully and informatively with a suitable caption.

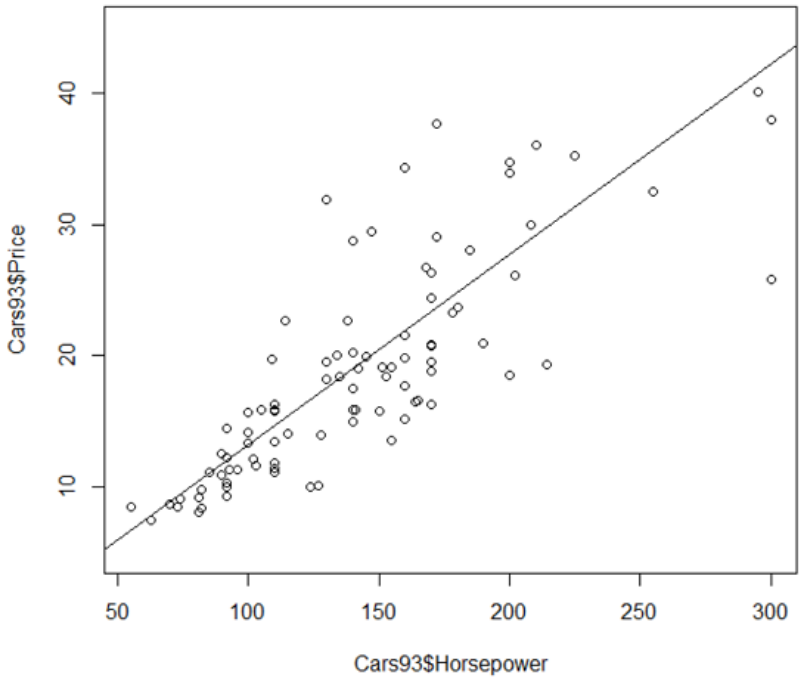
- Add labels to your axes on graphs.
- Show what the scale and unit of measurement is as well (e. raw numbers vs percentages).
- If using color or shading: Add an explanatory chart to show what each means.

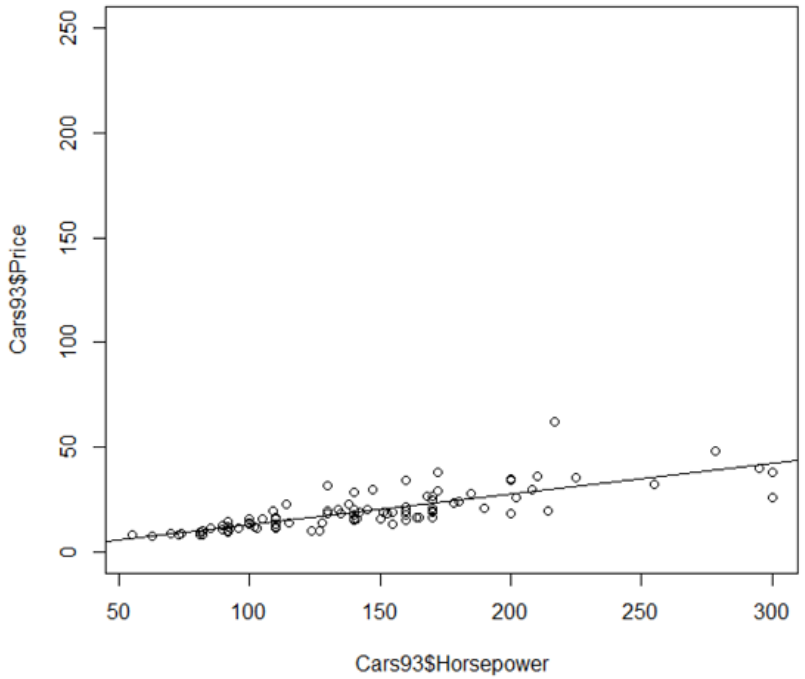
Selection of scale

- Choose the scale of the graph carefully.
- Most software defaults to using the observed range of a variable as the scale for the relevant axis.
- Sometimes this is OK, but it can also make the gradient of a trend line look more “impressive” than it actually is.

Compare the following…

Same data, different range of y-axis





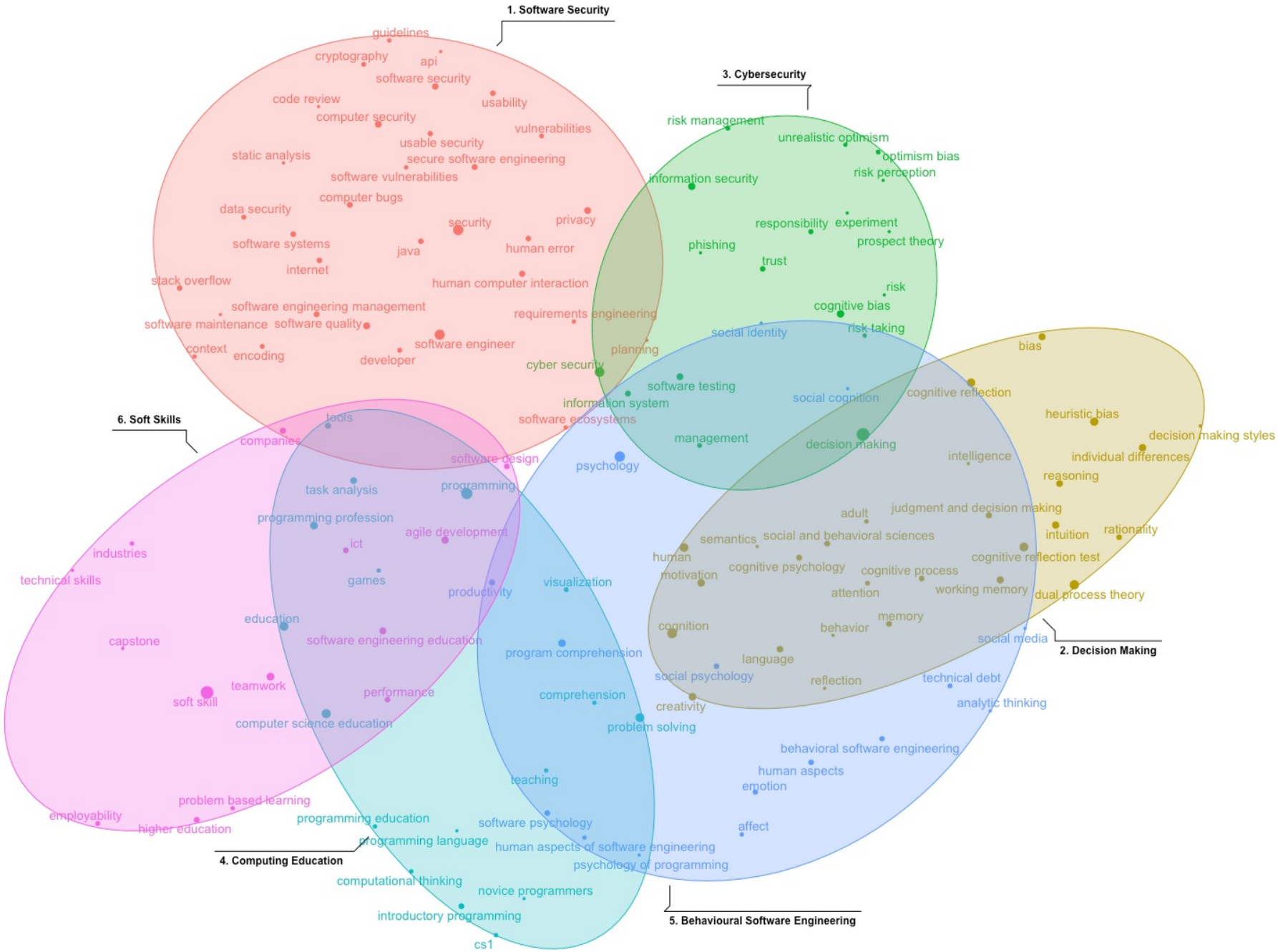
Avoid duplication

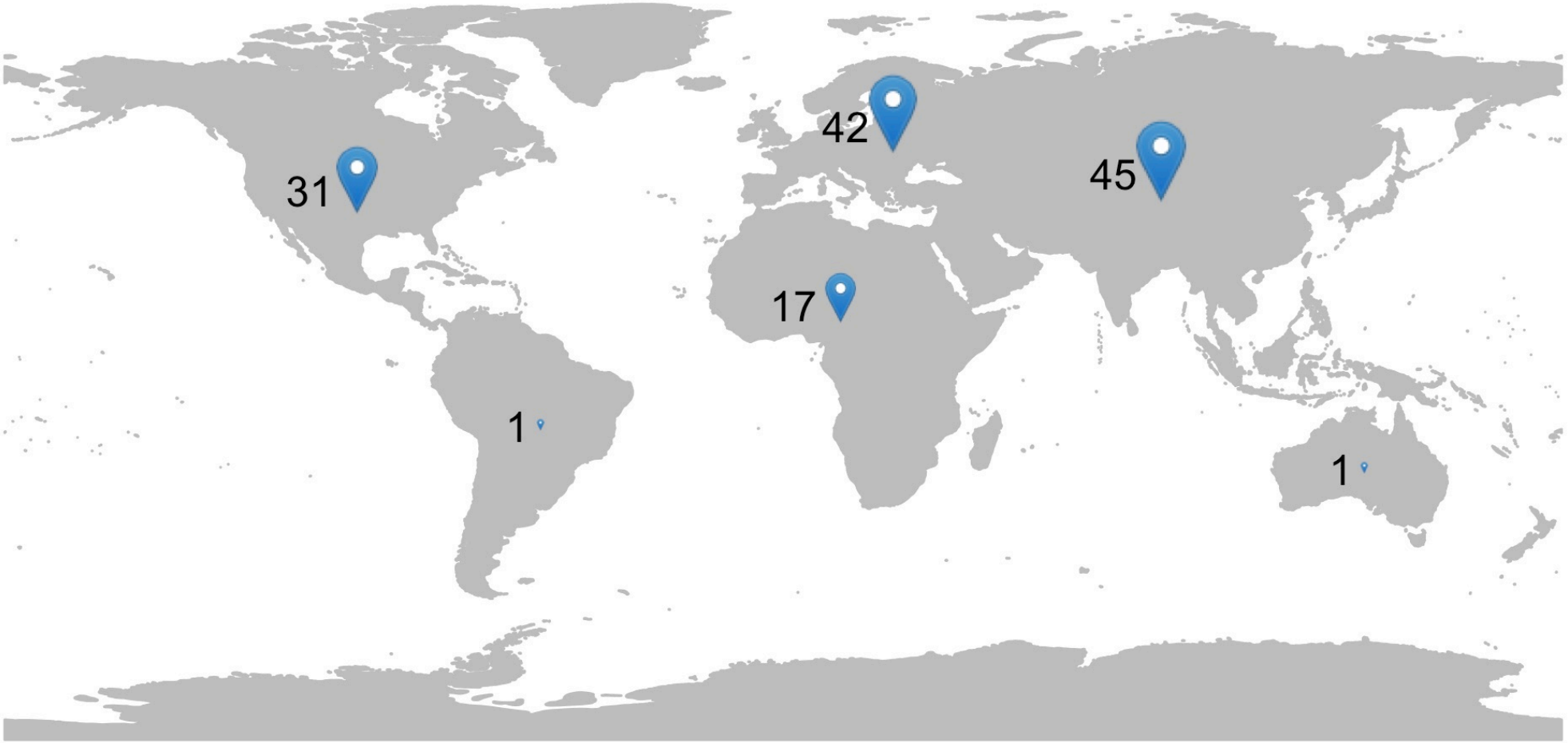
- Don’t use different kinds of plot to do the same “job” within the same article.
- Common mistake in undergraduate dissertations, e.g. students might use both bar charts and pie charts for the same data.
- Readers are interested in the data, not our virtuosity in using a graphics package

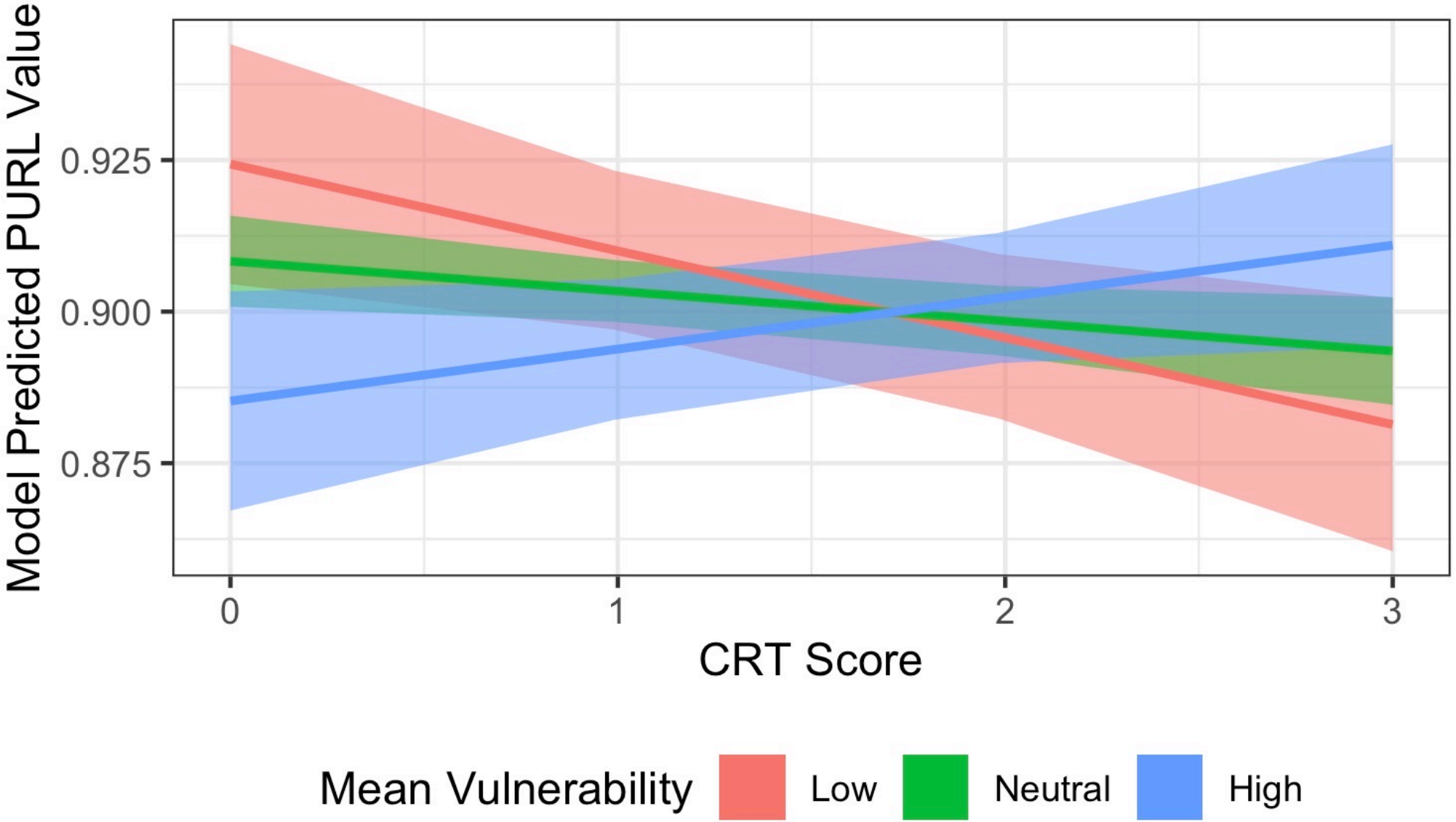
What do other people do?

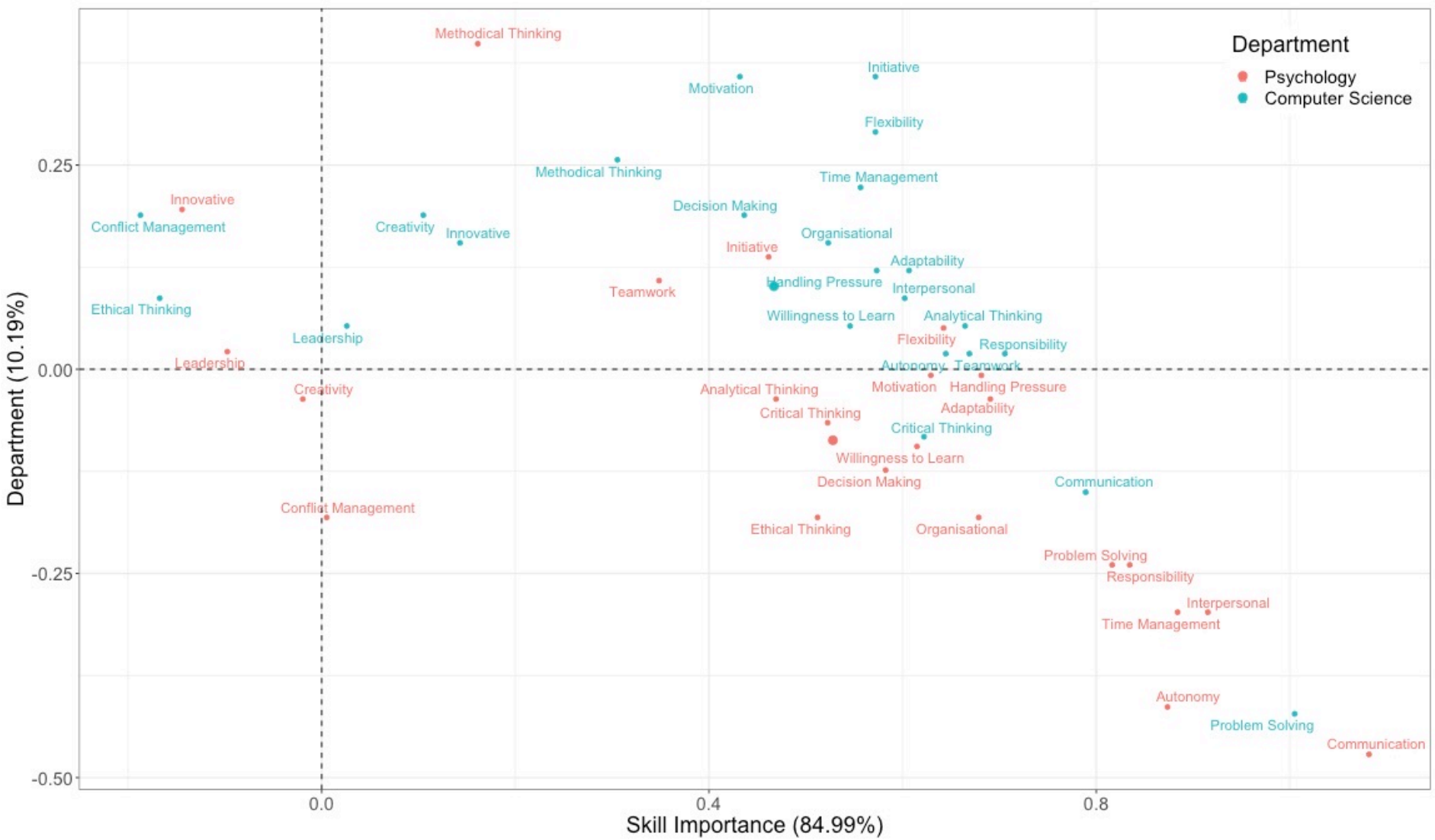
- You need to be aware of the reporting practices in your field.
- How do other researchers (those who publish now in leading peer-reviewed journals) display their data?
- Pay special attention to studies who use the same data or research instruments.

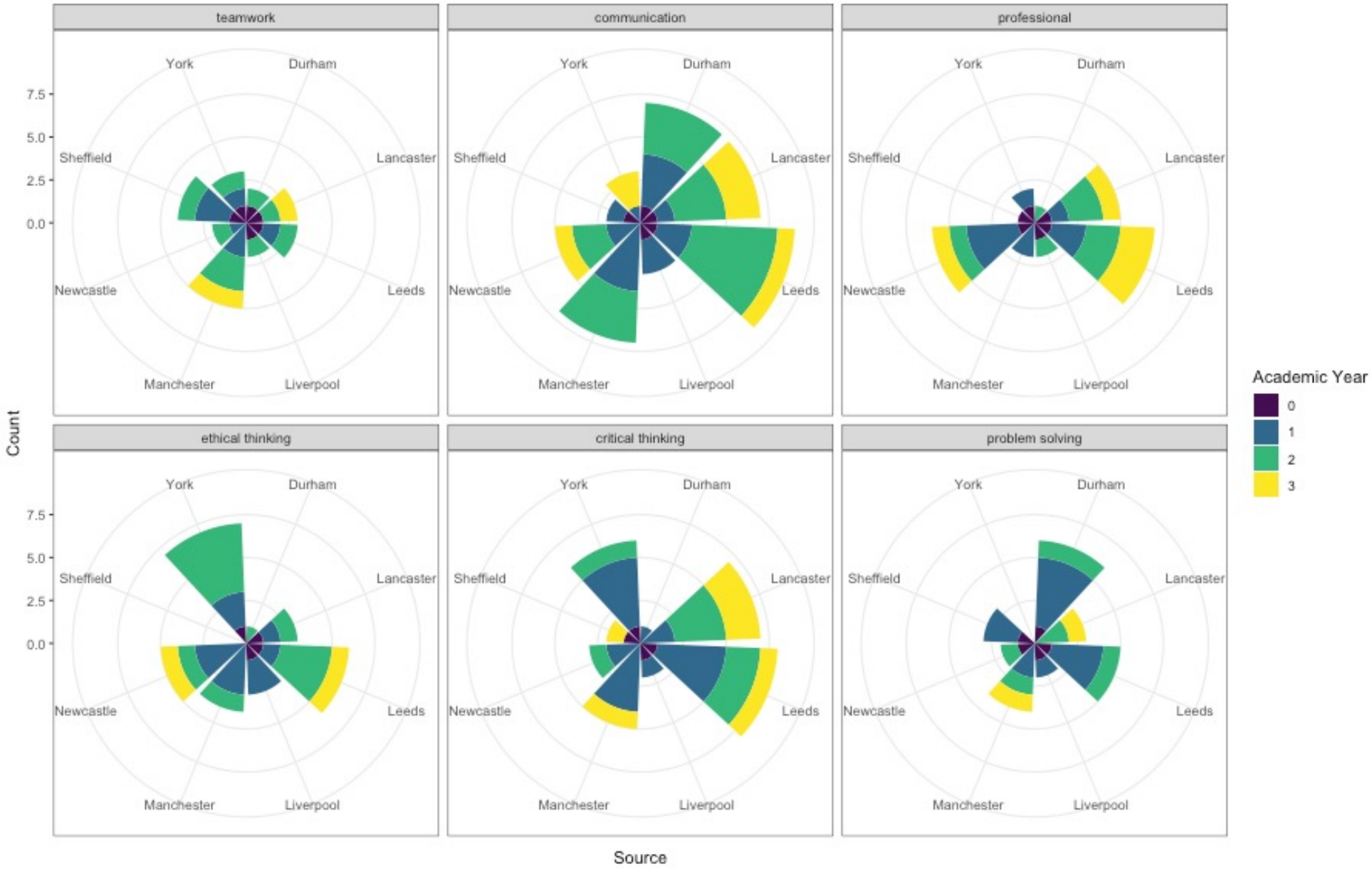
Examples: what can you plot?


1. Software Security
guidelines
cryptography
api
software security
code review
usability
computer security
vulnerabilities
usable security
static analysis
secure software engineering
software vulnerabilities
computer bugs
data security
security
privacy
software systems
java
human error
internet
stack overflow
human computer interaction
software engineering management
requirements engineering
software maintenance
software quality
software engineer
context
encoding
developer
3. Cybersecurity
risk management
unrealistic optimism
optimism bias
risk perception
information security
responsibility
experiment
prospect theory
phishing
trust
risk
cognitive bias
risk taking
social identity
planning
cyber security
software testing
information system
software ecosystems
management
psychology
software design
6. Soft Skills
tools
companies
task analysis
programming
programming profession
agile development
ict
industries
technical skills
games
productivity
education
capstone
software engineering education
teamwork
soft skill
performance
computer science education
visualization
program comprehension
comprehension
problem solving
teaching
problem based learning
employability
higher education
programming education
programming language
4. Computing Education
software psychology
human aspects of software engineering
psychology of programming
computational thinking
novice programmers
introductory programming
cs1
5. Behavioural Software Engineering
behavioral software engineering
human aspects
emotion
affect
social cognition
decision making
bias
cognitive reflection
heuristic bias
decision making styles
intelligence
individual differences
reasoning
adult
judgment and decision making
intuition
rationality
semantics
social and behavioral sciences
human
cognitive psychology
cognitive process
cognitive reflection test
motivation
attention
working memory
dual process theory
cognition
behavior
memory
social media
language
social psychology
reflection
technical debt
creativity
analytic thinking
2. Decision Making


42
31
45
17
1
1


Model Predicted PURL Value
0.925
0.900
0.875
0
1
2
3
CRT Score
Mean Vulnerability
Low
Neutral
High


Department
Psychology
Computer Science
Methodical Thinking
Initiative
Motivation
Flexibility
Methodical Thinking
Time Management
Innovative
Decision Making
Conflict Management
Creativity
Innovative
Organisational
Initiative
Adaptability
Teamwork
Handling Pressure
Interpersonal
Ethical Thinking
Willingness to Learn
Analytical Thinking
Leadership
Flexibility
Responsibility
Leadership
Autonomy
Teamwork
Creativity
Analytical Thinking
Motivation
Handling Pressure
Critical Thinking
Adaptability
Critical Thinking
Willingness to Learn
Decision Making
Communication
Conflict Management
Ethical Thinking
Organisational
Problem Solving
Responsibility
Interpersonal
Time Management
Autonomy
Problem Solving
Communication
0.25
0.00
-0.25
-0.50
0.0
0.4
0.8
Department (10.19%)
Skill Importance (84.99%)

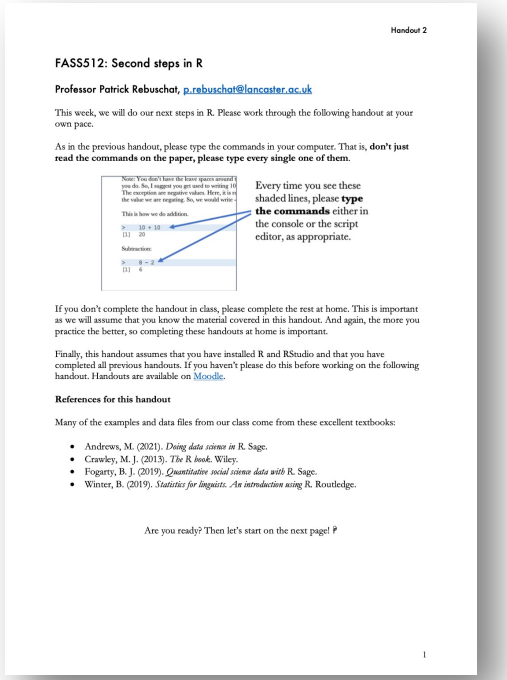

teamwork
communication
professional
ethical thinking
critical thinking
problem solving
York
Durham
Lancaster
Leeds
Liverpool
Manchester
Newcastle
Sheffield
7.5
5.0
2.5
0.0
Count
Source
Academic Year
0
1
2
3

Practical: Data visualization

Handout 4

Aim: Introduce you to two plotting systems in R, base R and `ggplot`.

`ggplot` is the most-widely used plotting system.



Handout 2

FASS512: Second steps in R

Professor Patrick Rebuschat, p.rebuschat@lancaster.ac.uk

This week, we will do our next steps in R. Please work through the following handout at your own pace.

As in the previous handout, please type the commands in your computer. That is, **don't just read the commands on the paper, please type every single one of them**.

Every time you see these shaded lines, please **type the commands** either in the console or the script editor, as appropriate.

If you don't complete the handout in class, please complete the rest at home. This is important as we will assume that you know the material covered in this handout. And again, the more you practice the better, so completing these handouts at home is important.

Finally, this handout assumes that you have installed R and RStudio and that you have completed all previous handouts. If you haven't please do this before working on the following handout. Handouts are available on Moodle.

References for this handout

Many of the examples and data files from our class come from these excellent textbooks:

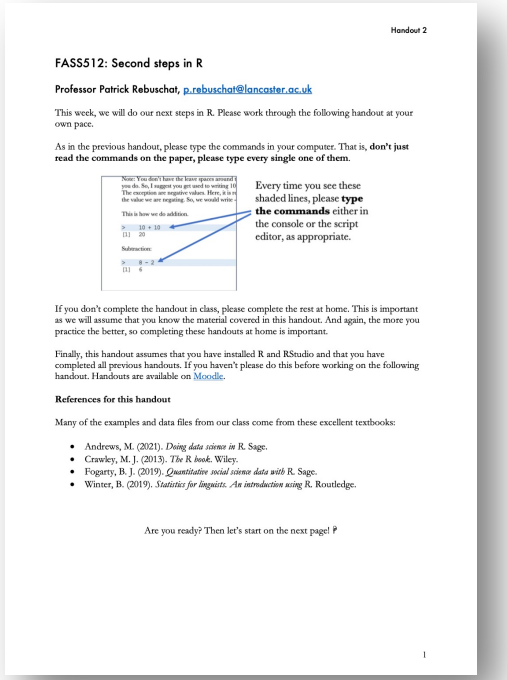
- Andrews, M. (2021). *Doing data science in R*. Sage.
- Crawley, M. J. (2013). *The R book*. Wiley.
- Fogarty, B. J. (2019). *Quantitative social science data with R*. Sage.
- Winter, B. (2019). *Statistics for linguists. An introduction using R*. Routledge.

Are you ready? Then let's start on the next page!

1

Handout 4: Web resources

- An excellent reference for ggplot
- A useful cheat sheet
- Gorgeous graphs in ggplot



Handout 2

FASS512: Second steps in R

Professor Patrick Rebuschat, p.rebuschat@lancaster.ac.uk

This week, we will do our next steps in R. Please work through the following handout at your own pace.

As in the previous handout, please type the commands in your computer. That is, **don't just read the commands on the paper, please type every single one of them**.

Every time you see these shaded lines, please **type the commands** either in the console or the script editor, as appropriate.

If you don't complete the handout in class, please complete the rest at home. This is important as we will assume that you know the material covered in this handout. And again, the more you practice the better, so completing these handouts at home is important.

Finally, this handout assumes that you have installed R and RStudio and that you have completed all previous handouts. If you haven't please do this before working on the following handout. Handouts are available on Moodle.

References for this handout

Many of the examples and data files from our class come from these excellent textbooks:

- Andrews, M. (2021). *Doing data science in R*. Sage.
- Crawley, M. J. (2013). *The R book*. Wiley.
- Fogarty, B. J. (2019). *Quantitative social science data with R*. Sage.
- Winter, B. (2019). *Statistics for linguists. An introduction using R*. Routledge.

Are you ready? Then let's start on the next page!

1

Data visualization with ggplot2 : : **CHEAT SHEET**

Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same components: a **data** set, a **coordinate system**, and **geoms**—visual marks that represent data points.

To display values, map variables in the data to visual properties of the geom (**aesthetics**) like **size**, **color**, and **x** and **y** locations.

Complete the template below to build a graph.

```
ggplot(data = <DATA>) +                                  required
<GEOM_FUNCTION>(mapping = aes(<MAPPINGS>),
  stat = <STAT>, position = <POSITION>) +                Not required, sensible defaults supplied
<COORDINATE_FUNCTION> +
<FACET_FUNCTION> +
<SCALE_FUNCTION> +
<THEME_FUNCTION>
```

ggplot(data = mpg, **aes**(x = cty, y = hwy)) Begins a plot that you finish by adding layers to. Add one geom function per layer.

last_plot() Returns the last plot.

ggsave("plot.png", width = 5, height = 5) Saves last plot as 5' x 5' file named "plot.png" in working directory. Matches file type to file extension.

Aes

Common aesthetic values.

color and **fill** - string ("red", "#RRGGBB")

linetype - integer or string (0 = "blank", 1 = "solid", 2 = "dashed", 3 = "dotted", 4 = "dotdash", 5 = "longdash", 6 = "twodash")

lineend - string ("round", "butt", or "square")

linejoin - string ("round", "mitre", or "bevel")

size - integer (line width in mm)

shape - integer/shape name or a single character ("a")

Geoms

Use a geom function to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

GRAPHICAL PRIMITIVES

a <- ggplot(economics, aes(date, unemploy))
b <- ggplot(seals, aes(x = long, y = lat))

a + geom_blank() and **a + expand_limits()** Ensure limits include values across all plots.

b + geom_curve(aes(yend = lat + 1, xend = long + 1), curvature = 1) - x, xend, y, yend, alpha, angle, color, curvature, linetype, size

a + geom_path(lineend = "butt", linejoin = "round", linemitre = 1) x, y, alpha, color, group, linetype, size

a + geom_polygon(aes(alpha = 50)) - x, y, alpha, color, fill, group, subgroup, linetype, size

b + geom_rect(aes(xmin = long, ymin = lat, xmax = long + 1, ymax = lat + 1)) - xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size

a + geom_ribbon(aes(ymin = unemploy - 900, ymax = unemploy + 900)) - x, ymax, ymin, alpha, color, fill, group, linetype, size

LINE SEGMENTS

common aesthetics: x, y, alpha, color, linetype, size

b + geom_abline(aes(intercept = 0, slope = 1))
b + geom_hline(aes(yintercept = lat))
b + geom_vline(aes(xintercept = long))

b + geom_segment(aes(yend = lat + 1, xend = long + 1))
b + geom_spoke(aes(angle = 1:1155, radius = 1))

ONE VARIABLE continuous

c <- ggplot(mpg, aes(hwy)); c2 <- ggplot(mpg)

c + geom_area(stat = "bin") x, y, alpha, color, fill, linetype, size

c + geom_density(kernel = "gaussian") x, y, alpha, color, fill, group, linetype, size, weight

c + geom_dotplot() x, y, alpha, color, fill

c + geom_freqpoly() x, y, alpha, color, group, linetype, size

c + geom_histogram(binwidth = 5) x, y, alpha, color, fill, linetype, size, weight

c2 + geom_qq(aes(sample = hwy)) x, y, alpha, color, fill, linetype, size, weight

discrete

d <- ggplot(mpg, aes(fl))

d + geom_bar() x, alpha, color, fill, linetype, size, weight

TWO VARIABLES both continuous

e <- ggplot(mpg, aes(cty, hwy))

e + geom_label(aes(label = cty), nudge_x = 1, nudge_y = 1) - x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

e + geom_point() x, y, alpha, color, fill, shape, size, stroke

e + geom_quantile() x, y, alpha, color, group, linetype, size, weight

e + geom_rug(sides = "bl") x, y, alpha, color, linetype, size

e + geom_smooth(method = lm) x, y, alpha, color, fill, group, linetype, size, weight

e + geom_text(aes(label = cty), nudge_x = 1, nudge_y = 1) - x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

one discrete, one continuous

f <- ggplot(mpg, aes(class, hwy))

f + geom_col() x, y, alpha, color, fill, group, linetype, size

f + geom_boxplot() x, y, lower, middle, upper, ymax, ymin, alpha, color, fill, group, linetype, shape, size, weight

f + geom_dotplot(binaxis = "y", stackdir = "center") x, y, alpha, color, fill, group

f + geom_violin(scale = "area") x, y, alpha, color, fill, group, linetype, size, weight

both discrete

g <- ggplot(diamonds, aes(cut, color))

g + geom_count() x, y, alpha, color, fill, shape, size, stroke

e + geom_jitter(height = 2, width = 2) x, y, alpha, color, fill, shape, size

continuous bivariate distribution

h <- ggplot(diamonds, aes(carat, price))

h + geom_bin2d(binwidth = c(0.25, 500)) x, y, alpha, color, fill, linetype, size, weight

h + geom_density_2d() x, y, alpha, color, group, linetype, size

h + geom_hex() x, y, alpha, color, fill, size

continuous function

i <- ggplot(economics, aes(date, unemploy))

i + geom_area() x, y, alpha, color, fill, linetype, size

i + geom_line() x, y, alpha, color, group, linetype, size

i + geom_step(direction = "hv") x, y, alpha, color, group, linetype, size

visualizing error

df <- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2)
j <- ggplot(df, aes(grp, fit, ymin = fit - se, ymax = fit + se))

j + geom_crossbar(fatten = 2) - x, y, ymax, ymin, alpha, color, fill, group, linetype, size

j + geom_errorbar() - x, ymax, ymin, alpha, color, group, linetype, size, width Also **geom_errorbarh()**.

j + geom_linerange() x, ymin, ymax, alpha, color, group, linetype, size

j + geom_pointrange() - x, y, ymin, ymax, alpha, color, fill, group, linetype, shape, size

maps

data <- data.frame(murder = USArrests$Murder, state = tolower(rownames(USArrests)))
map <- map_data("state")
k <- ggplot(data, aes(fill = murder))

k + geom_map(aes(map_id = state), map = map) **+ expand_limits**(x = map$long, y = map$lat) map_id, alpha, color, fill, linetype, size

THREE VARIABLES

seals$z <- with(seals, sqrt(delta_long^2 + delta_lat^2)); l <- ggplot(seals, aes(long, lat))

l + geom_contour(aes(z = z)) x, y, z, alpha, color, group, linetype, size, weight

l + geom_contour_filled(aes(fill = z)) x, y, alpha, color, fill, group, linetype, size, subgroup

l + geom_raster(aes(fill = z), hjust = 0.5, vjust = 0.5, interpolate = FALSE) x, y, alpha, fill

l + geom_tile(aes(fill = z)) x, y, alpha, color, fill, linetype, size, width

Questions?

- I will be walking around while you work through the worksheet