Exercise 2 | Motivation

Max Pellert

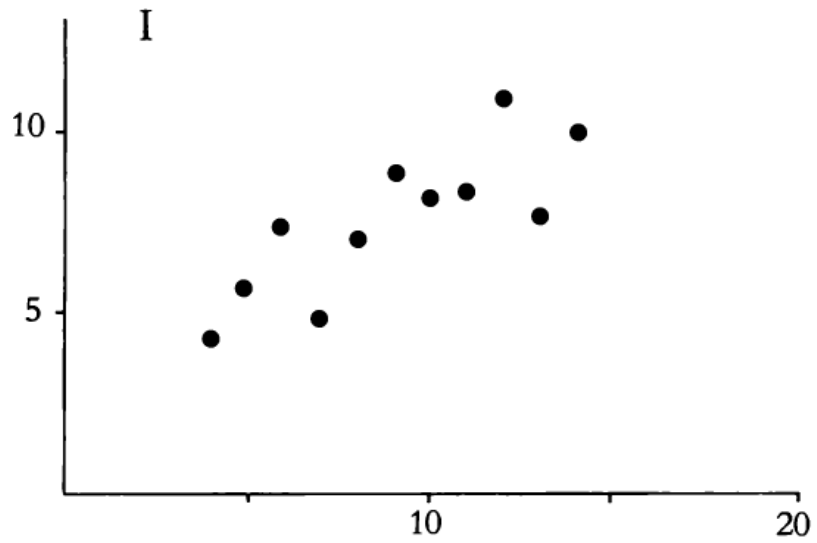
IS 616: Large Scale Data Analysis and Visualization

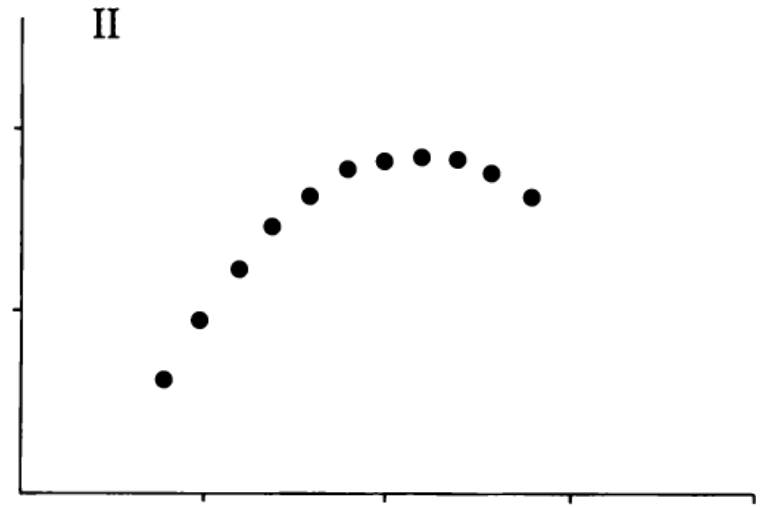

0:00

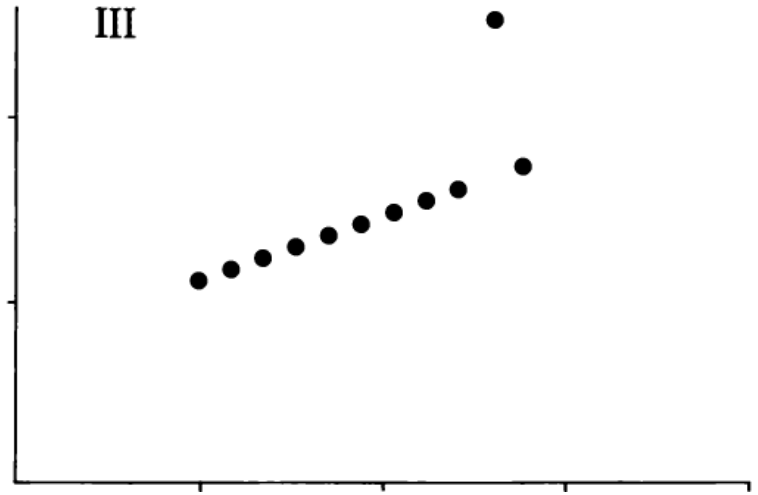
I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

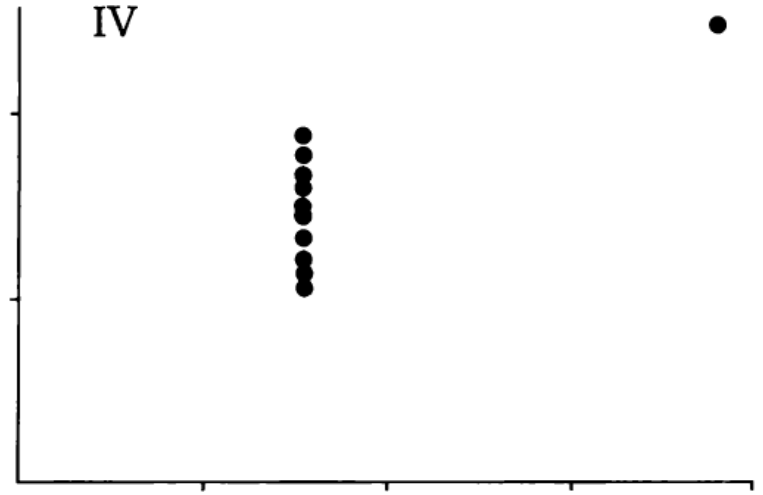
→

$N = 11$
mean of X's $= 9.0$
mean of Y's $= 7.5$
equation of regression line: $Y = 3 + 0.5X$
standard error of estimate of slope $= 0.118$
$t = 4.24$
sum of squares $X - \overline{X} = 110.0$
regression sum of squares $= 27.50$
residual sum of squares of $Y = 13.75$
correlation coefficient $= .82$
$r^2 = .67$


I
10
5
10
20


II


III


IV

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.

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*

Anscombe, F. J. (1973). Graphs in Statistical Analysis. The American Statistician, 27(1), 17–21. https://doi.org/10.1080/00031305.1973.10478966

"Anscombes quartett" shows that sometimes (or better always) you need to visualize!

Never forget about nonlinearity! (e.g. no relationship according to measures such Pearson correlation but possibly many functional relationships that are not linear)

README.md

Awesome data journalism

A partial, curated list of publicly available, free/open source and open access resources for learning and doing data journalism.

This repository builds on lists and collections of resources from the first and second editions of the open access *Data Journalism Handbook* and ongoing research on data journalism practices.

It is used and updated as part of open educational resources for the data journalism MA module at King's College London.

Suggestions for open access resources or links to add are most welcome. The repository for this page is here.

https://github.com/jwyg/awesome-data-journalism

Data Journalism

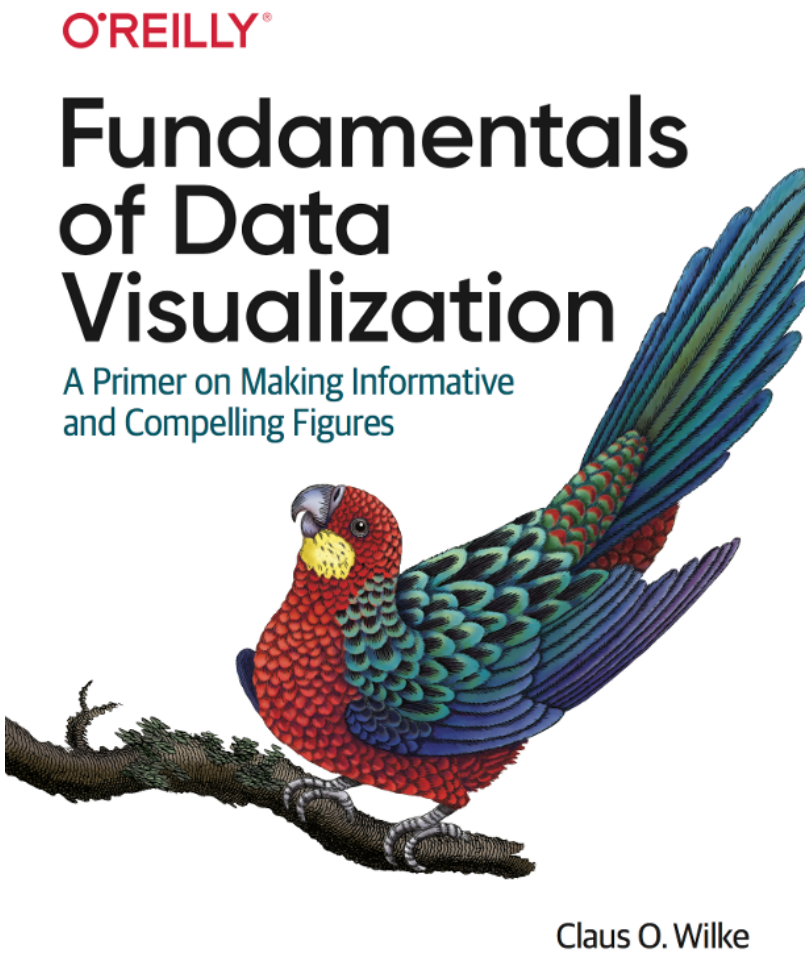
A weekly series: https://gijn.org/series/top-10-data-journalism-links/

https://www.nytimes.com/section/upshot

https://www.spiegel.de/thema/daten/

https://www.nzz.ch/visuals

…



https://clauswilke.com/dataviz/

Data Visualization Books

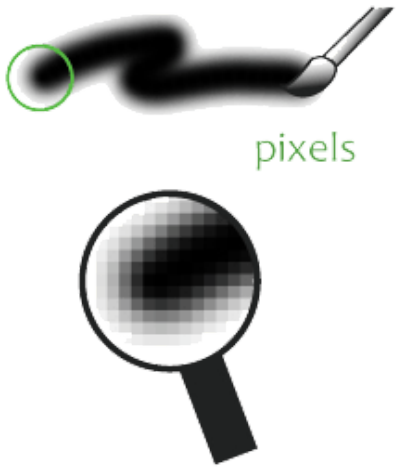
Roger D. Peng & Elizabeth Matsui: The Art of Data Science (https://bookdown.org/rdpeng/artofdatascience/)

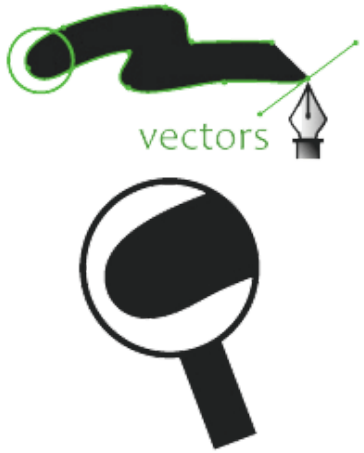
Kieran Healy: Data Visualization (https://socviz.co/index.html)

Winston Chang: ggplot2 cookbook (http://www.cookbook-r.com/Graphs/)

Jake VanderPlas: Python Data Science Handbook (https://jakevdp.github.io/PythonDataScienceHandbook/)

BBC Data Journalism team (https://medium.com/bbc-visual-and-data-journalism/how-the-bbc-visual-and-data-journalism-team-works-with-graphics-in-r-ed0b35693535)


pixels


vectors

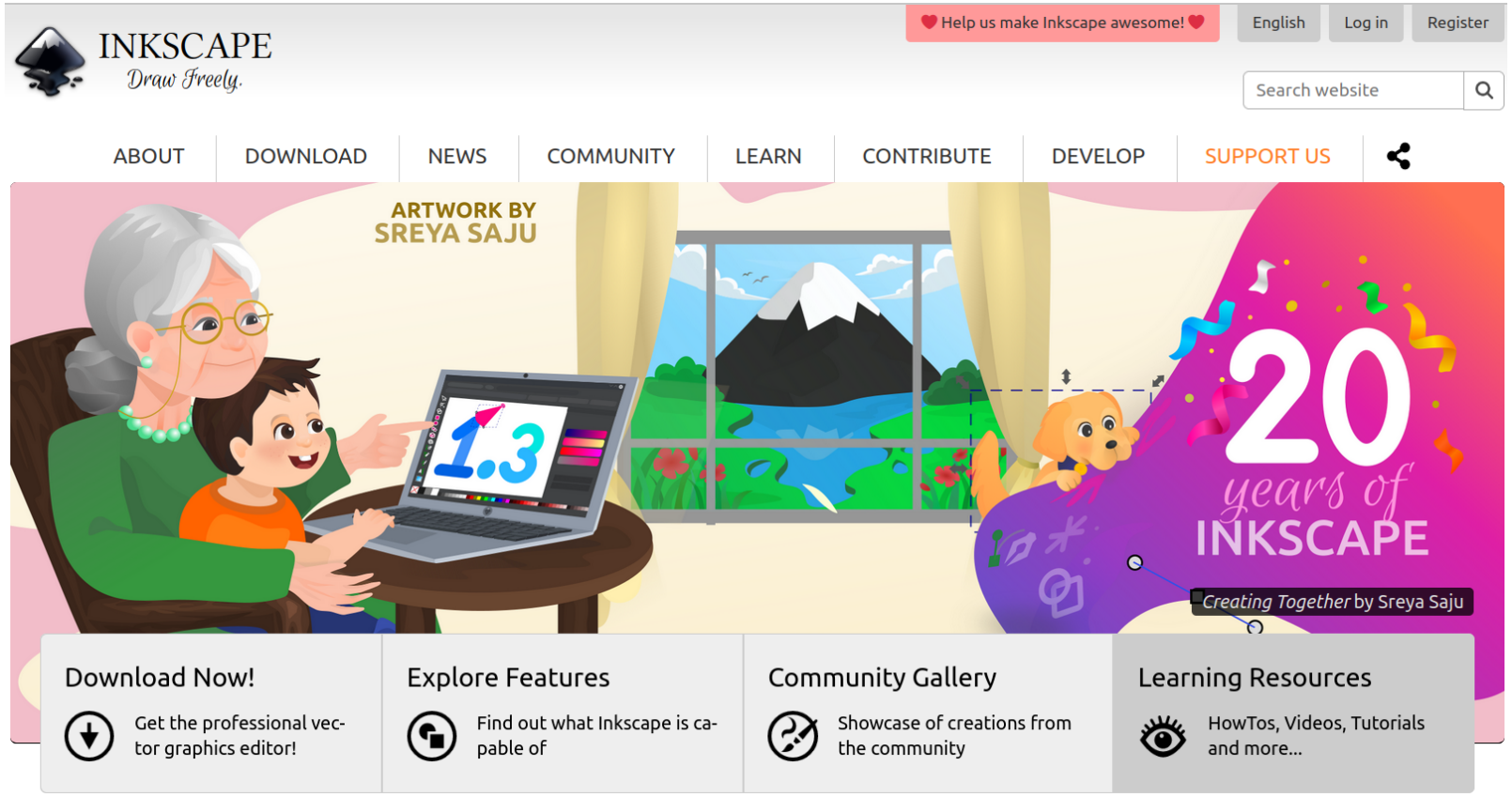
Vector vs bitmap (pixel-based) graphics

General rule: always save your plots as vector graphics!

Usually it works best if you save them as PDF files

Exception to the rule may be plots with a very large number of elements (for example scatterplots of many rows): it can make sense to rasterize those elements to get smaller files that can be handled more easily

I you have to use bitmap files (png, jpeg, …) make sure to save them with enough DPI (>300)



Users

A powerful, free design tool

Whether you are an illustrator, designer, web designer or just someone who needs to create some vector imagery, Inkscape is for you!

- Flexible drawing tools
- Broad file format compatibility
- Powerful text tool
- Bezier and spiro curves

Want to find out more about how Inkscape can help you? Look at the full set of features or try it!

Inkscape

"Inkscape is professional quality vector graphics software" that was developed by many different people over time

This can lead to idiosyncracies that are often the result of historical artefacts

A very steep learning curve in the beginning

But you don't need to master all of it

It's very powerful for often occurring small tasks, for example moving a legend to the right place

Playing around with Inkscape can lead to new insights in how to transmit most information using the "least ink"

Use the remaining time to

Think of visualizations that made an impression on you (from scientific papers, data journalism, blog posts, …)

Select one really good example and keep a note on it

This example should serve as a motivating example that you return to during the course (for example to judge it according to information design principles that we will discuss)

Use the remaining time to

Write down you personal motivation for the course in a few sentences

Find a data set with more than 10 000 and less than 100 000 rows that you want to study

Prepare reasons why exactly that data and what you expect to find in it, outline research questions