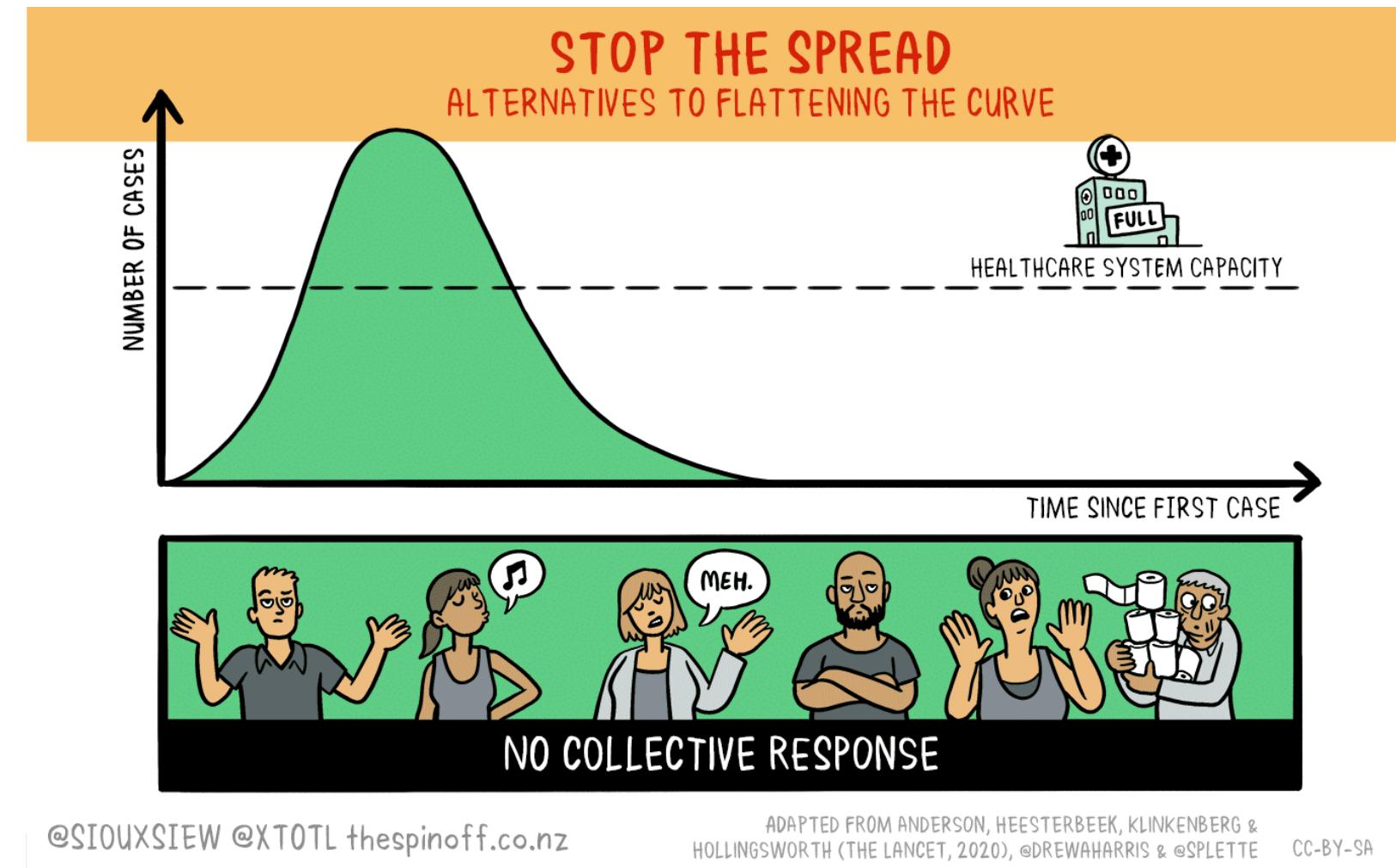


Introduction to *Urban* Data Science

EDA &
Visualisation
(EPA1316)
Lecture 5

Trivik Verma



Last Time

- Descriptive Statistics
- Data Transformations

Today

- History of Visualisations
- Exploratory Data Analysis
- Types of Visualisations
- Effective Visualisation

History

*“Data graphics **visually display measured quantities** by means of the **combined use** of points, lines, a coordinate system, numbers, symbols, words, shading, and color.”*

The Visual Display of Quantitative Information. Edward R. Tufte.

Tufte (1983)

“The most extensive data maps place millions of bits of information on a single page before our eyes. No other method for the display of statistical information is so powerful”

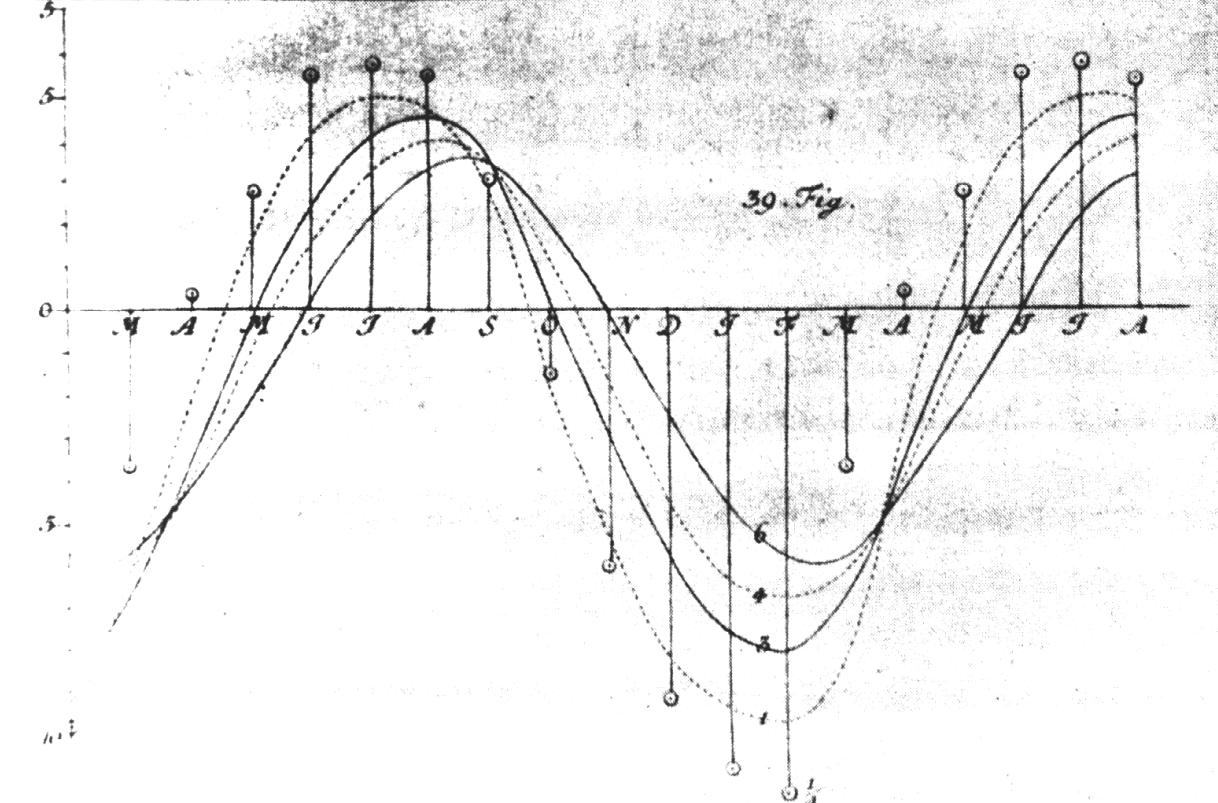
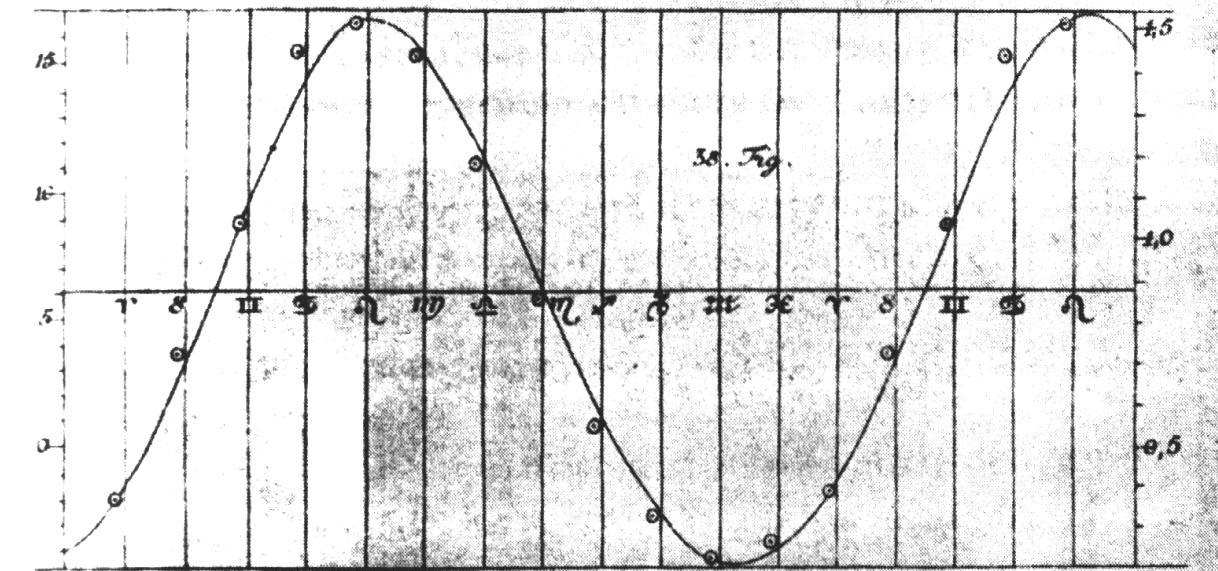
A bit of history

Maps → Data Maps (XVIIth.C.) → Time series (1786) → Scatter plots

- Surprisingly recent: 1750-1800 approx. (much later than many other advances in math and stats!)
- **William Playfair's “linear arithmetic”:**
encode/replace numbers in tables into visual representations.
- Other relevant names throughout history:
Lambert, Minard, Marey

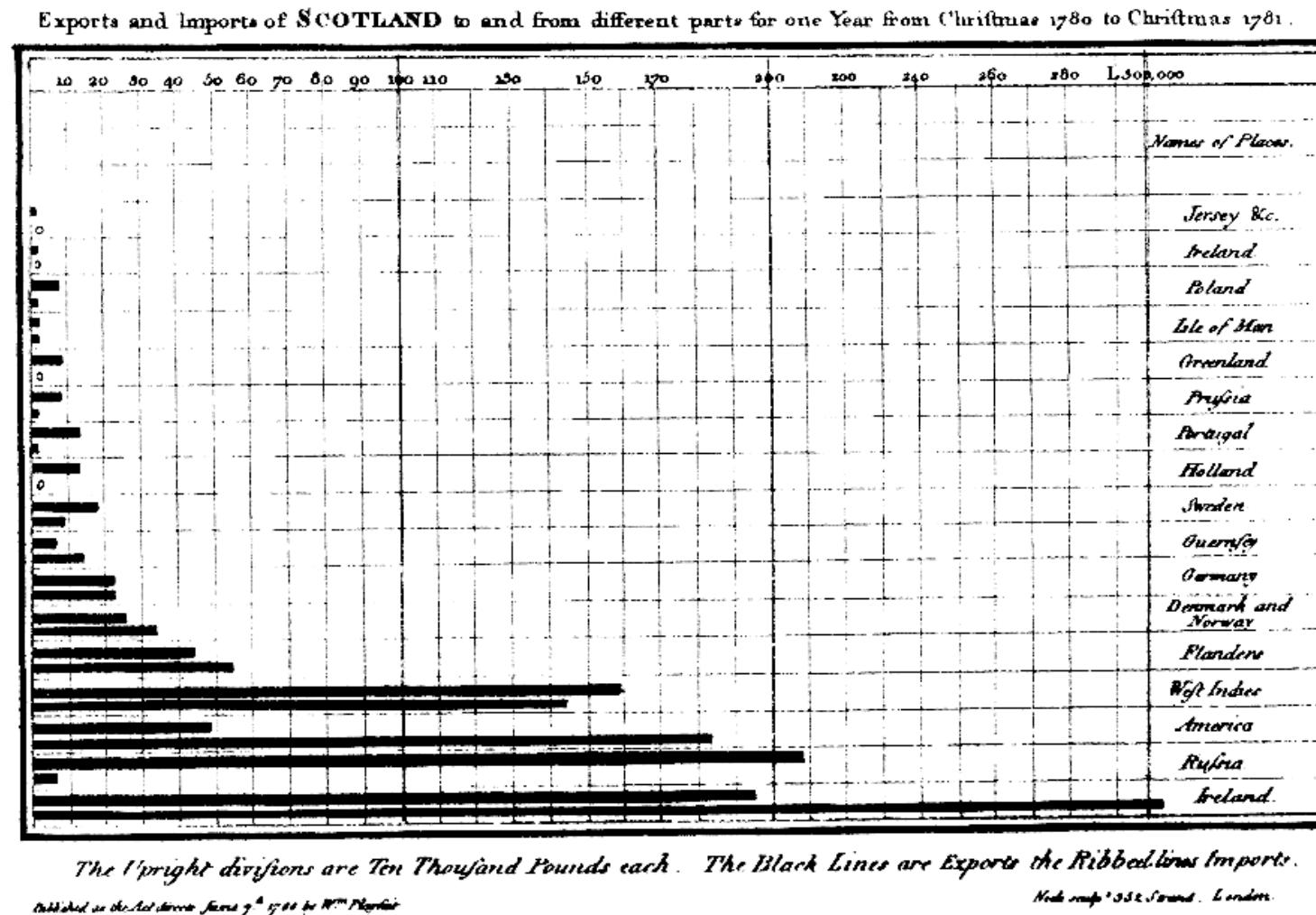
Historical Examples

[Source] XVIIIth. Cent. - *Pyrometrie* by J. H Lambert



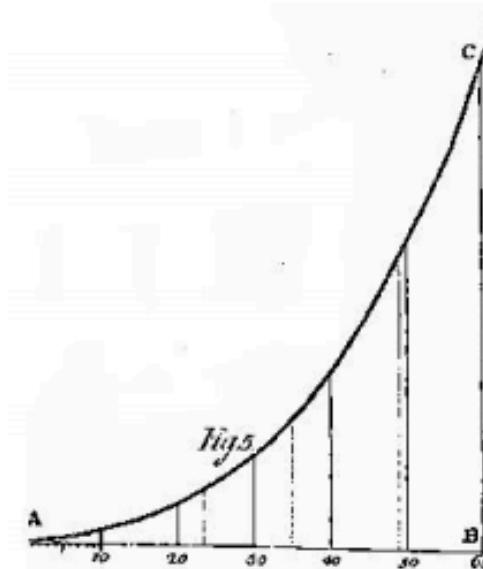
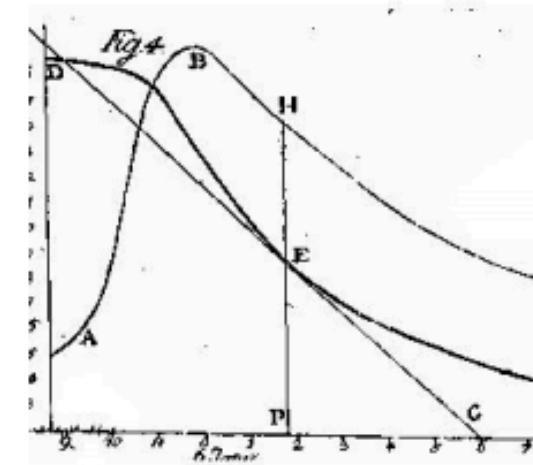
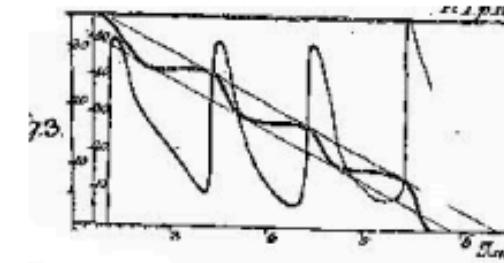
Historical Examples

[[Source](#)] Playfair's bar chart in The Commercial and Political Atlas (1786)



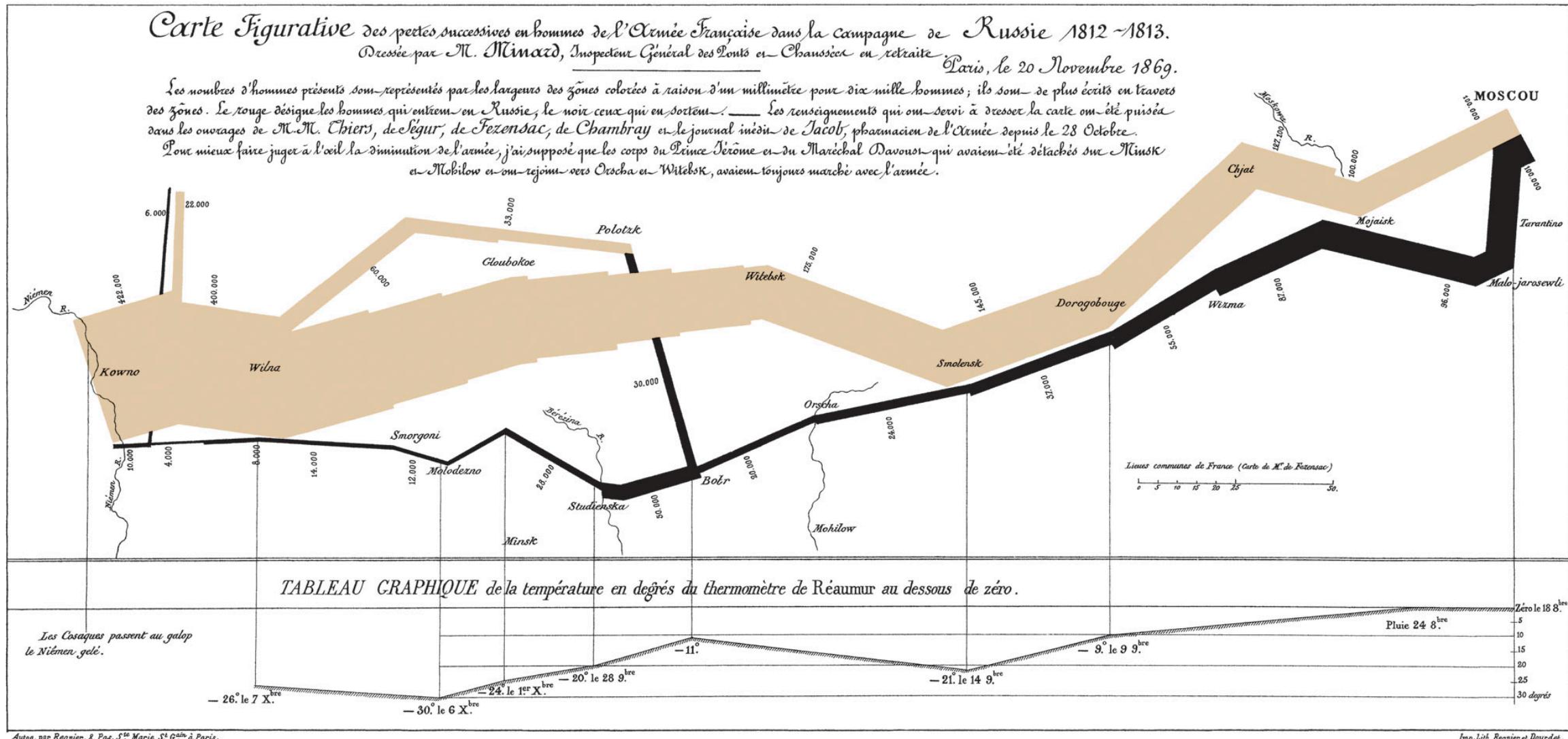
Historical Examples

[[Source](#)] Lambert - Evaporation rate against temperature, 1769



Historical Examples

[Source] Minard - Napoleon army map (XIXth. Cent.)



Challenges with Data

- The size of datasets from 10 years ago that were difficult to visualize can now be handled in real time on ***high-tech hardware***.
But the datasets of today have simply grown to be just as problematic.
- Datasets are getting larger as ***gathering resolution improves***.
- Datasets are getting larger as ***compute resources grow*** allowing higher resolution simulations.

Visualisation Goals

Analyse (Exploratory)

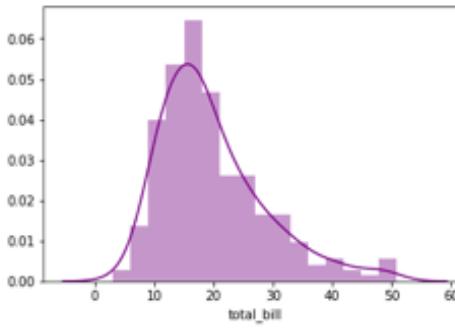
- Explore the data
- Assess a situation
- Identify hidden patterns and trends
- Formulate/test hypothesis
- Decide what to do next in analysis/modelling

Communicate (Explanatory)

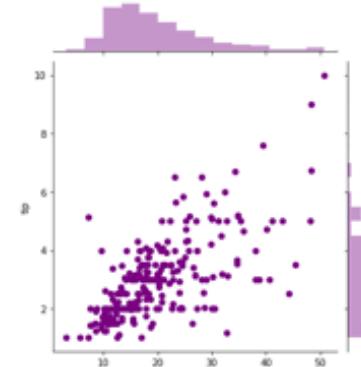
- Present information and ideas succinctly
- Explain and inform
- Provide evidence and support
- Influence and persuade

Visualisation Goals

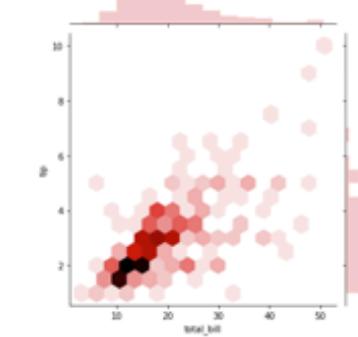
Analyse (Exploratory)



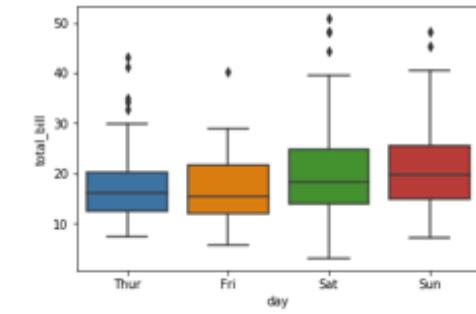
distplot



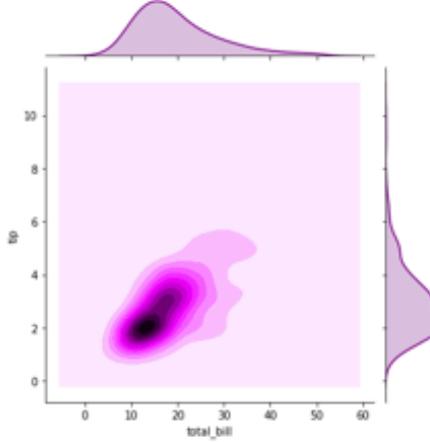
Jointplot



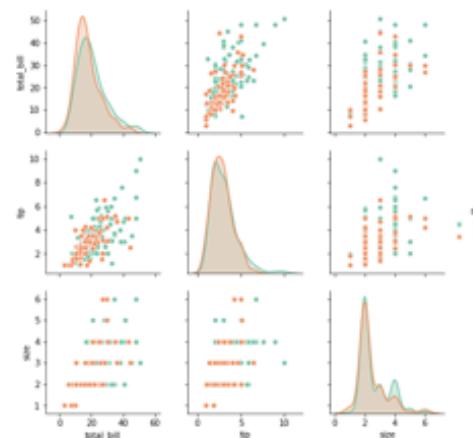
Hexplots



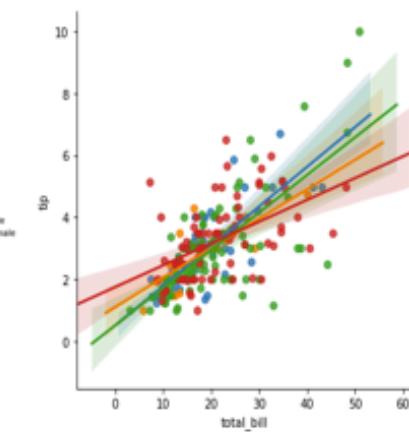
Boxplots



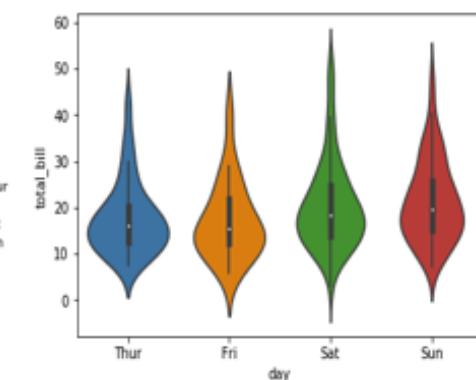
KDE Plot



Pair Plots



LM Plots



Violin Plots

Visualisation Goals

Communicate (Explanatory)

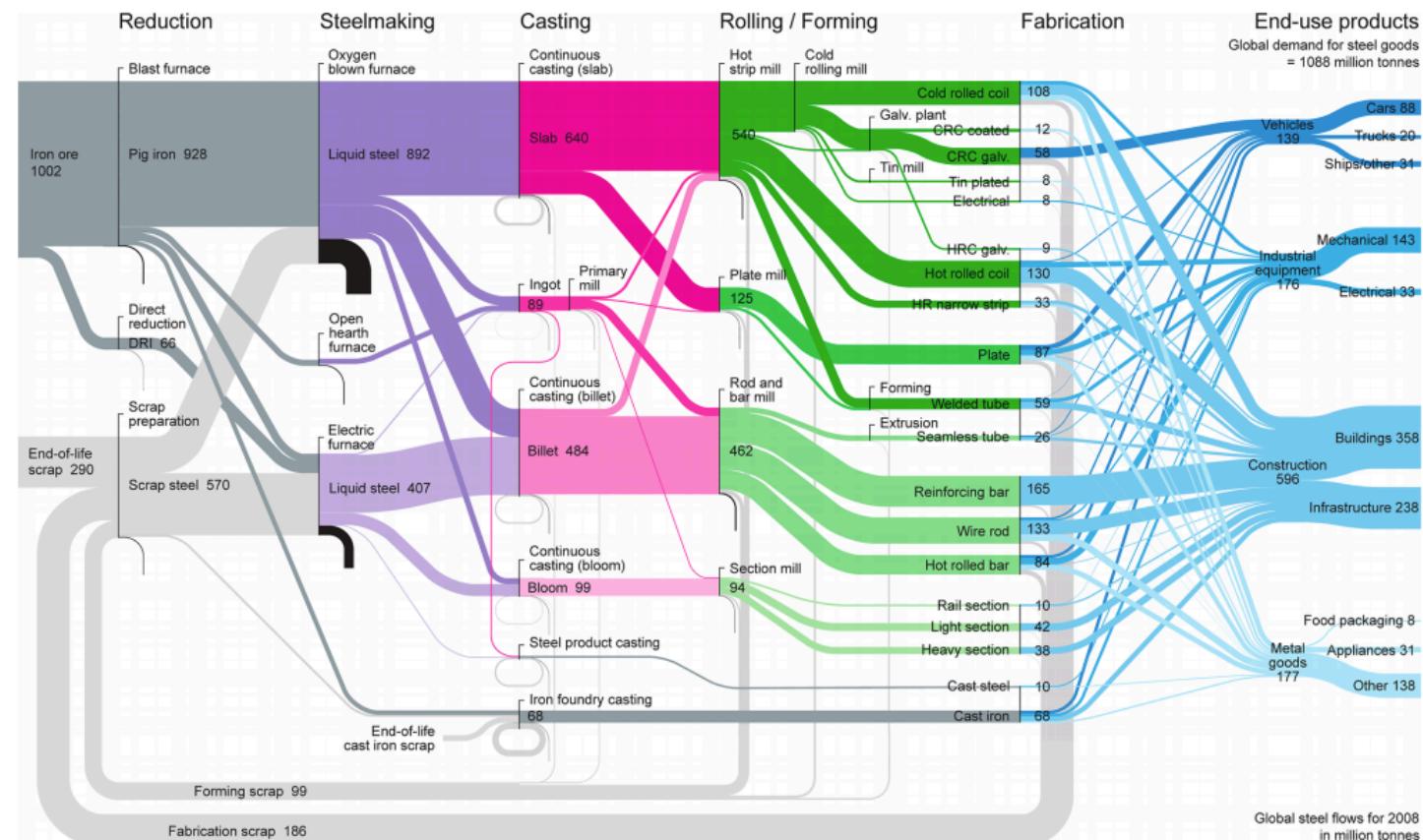


Figure 1. Global flow of steel from liquid metal to end-use good.

Tables vs Graphs

(56)

The Reader, that he hath found, that the Apertures, which Optick-Glasses can bear with distinctnes, are in about a subduplicate proportion to their Lengths; whereof he tells us he intends to give the reason and demonstration in his Dioptrick, which he is now writing, and intends to finish, as soon as his Health will permit. In the mean time, he presents the Reader with a Table of such Apertures; which is here exhibited to the Consideration of the Ingenious; there being of this French Book but one Copy, that is known, in England.

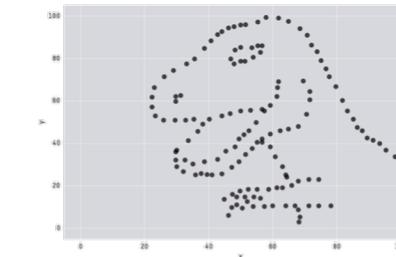
A TABLE of the Apertures of Object-Glasses.
The Points put to some of these Numbers denote Fractions.

Lengths of Glasses, in Feet, Inches, & Parts.	For exceeding Five feet.			For ordinary Lengths of Glasses.			For exceeding Five feet.			For ordinary Lengths, in Feet, Inches, & Parts.		
	Feet, & even Inches.	Feet, even Inches.	Feet, even Inches.	Feet, & even Inches.	Feet, even Inches.	Feet, & even Inches.	Feet, & even Inches.	Feet, even Inches.	Feet, even Inches.	Feet, & even Inches.	Feet, even Inches.	Feet, & even Inches.
4	4.	4	3 25	3	4 2	10 2	4 1					
5	5	5	4 30	3	8 3	2 2	7					
6	7	6	5 35	4	0 3	4 2	10					
7	8	7	6 40	4	3 3	7 3	1					
8												
9	6	8	7 45	4	6 3	10 3	2					
10	11	10	8 50	4	9 4	0 3	4					
11	0	11	9 55	5	0 4	3 3	6					
12	0 1	12	10 60	5	2 4	6 3	8					
13	6 1	2 1	11 65	5	4 4	8 3	10					
14	0 1	4 1	0 70	5	7 4	10 4						
15	6 1	5 1	.75	5	9 5	0 4	2					
16	0 1	6 1	1.80	5	11 5	2 4	5					
17	1	7 1	2 90	6	4 5	6 4	7					
18	1	9 1	6 10	6	8 5	9 4	10					
19	1	10 1	8 1	7	5 6	5 5	3					
20	1	11 1	9 1	7	0 7	0 5	12					
21	2	1 1	10 1	6 200	9	6 8	0 6	9				
22	2	4 2	0 1	8 250	10	6 9	2 7	8				
23	2	6 2	2 1	9 300	11	6 10	0 8	5				
24	2	8 2	4 1	11 350	12	6 10	9 9	6				
25	2	10 2	6 2	13 400	13	4 11	6 9	8				
26	3	0 2	7 2	2 1								

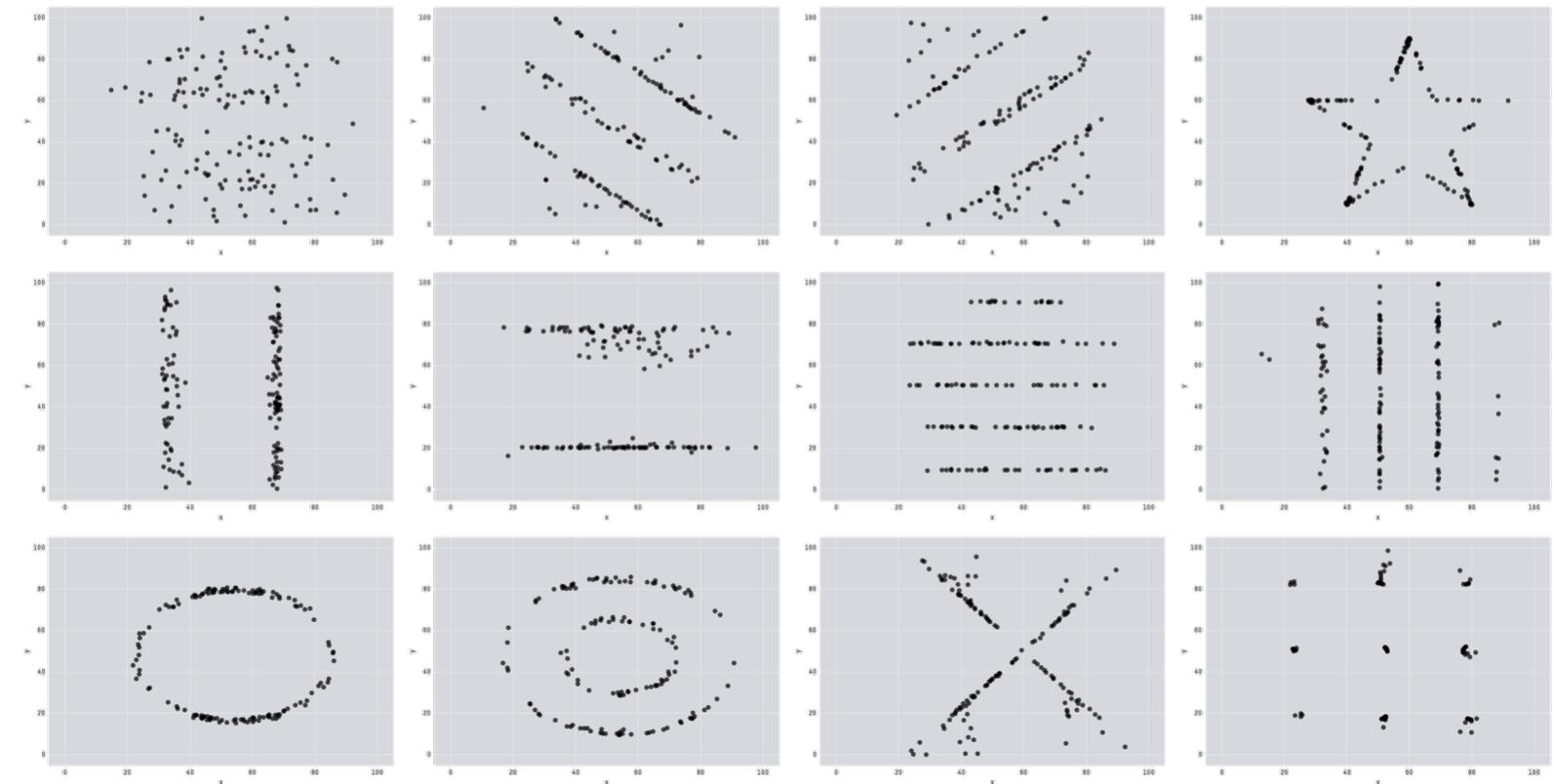
- Tables are generally best if you want to be able to look up specific information or if the values must be reported precisely
- By **encoding information visually**, they allow to present **large amounts of numbers in a meaningful way**
- Graphics are best for illustrating trends and making comparisons
- If well made, visualizations provide leads into the **processes** underlying the graphic
- Modern data graphics can do much more than simply substitute for small statistical tables
- Graphics are instruments for reasoning about quantitative information

Graphics reveal data

Anscombe's quartet



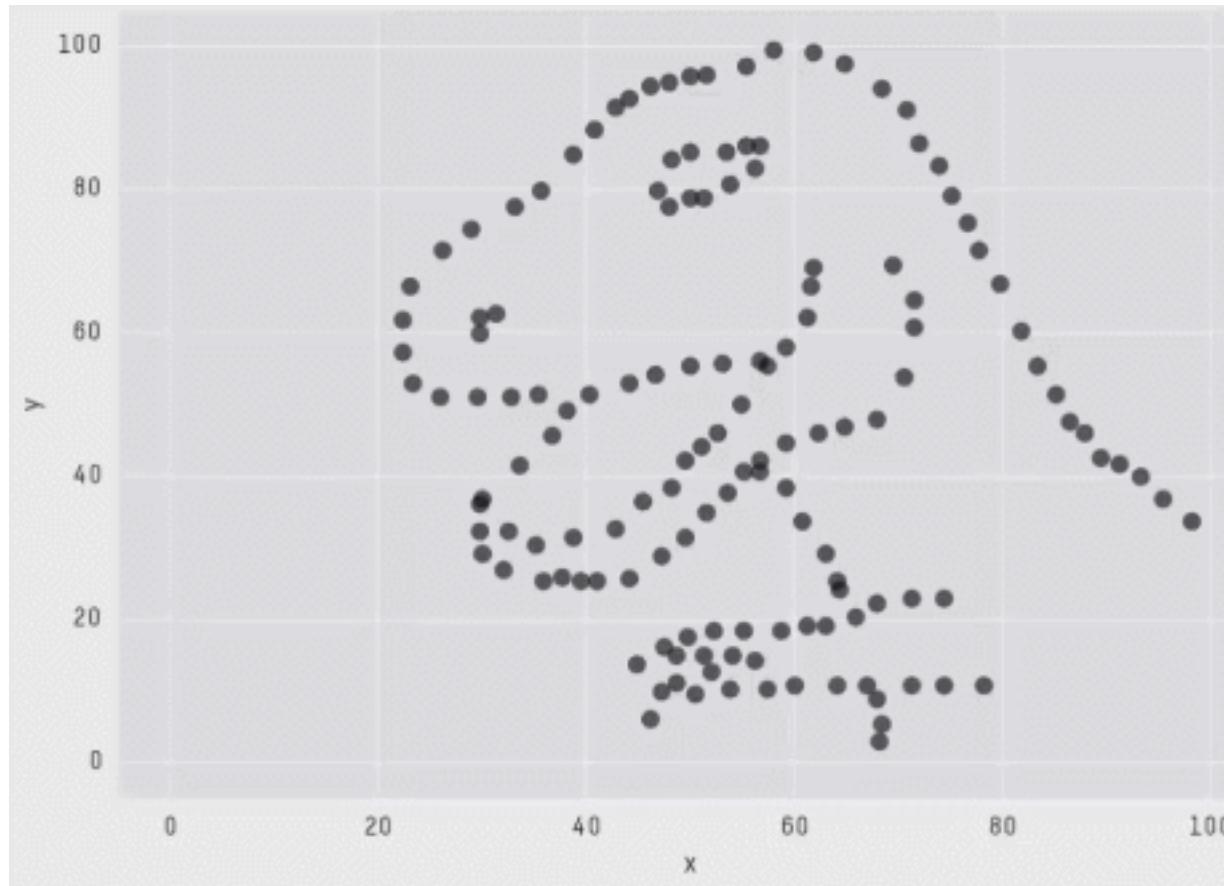
X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06



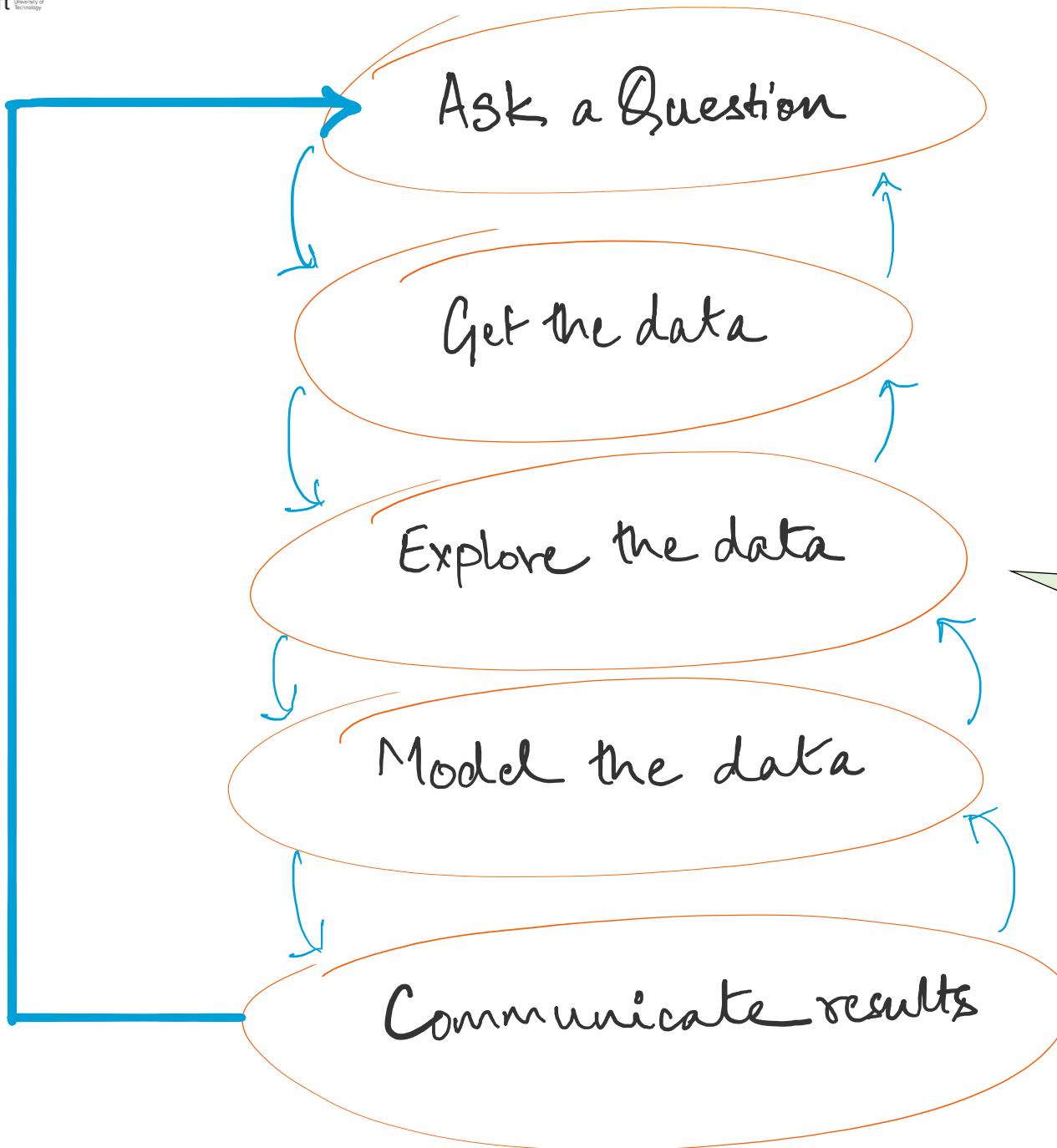
Graphics reveal data

Anscombe's quartet

Summary statistics clearly don't tell the story of how they differ. But a picture can be worth a thousand words.



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

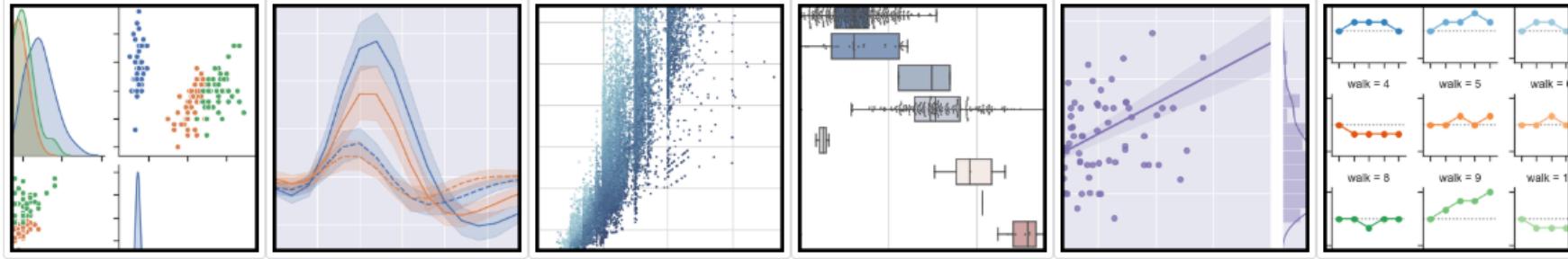


Exploratory Data Analysis (EDA)

To convey information through graphical representations of data

seaborn 0.10.0 [Gallery](#) [Tutorial](#) [API](#) [Site ▾](#) [Page ▾](#) [Search](#)

seaborn: statistical data visualization



Seaborn is a Python data visualization library based on [matplotlib](#). It provides a high-level interface for drawing attractive and informative statistical graphics.

For a brief introduction to the ideas behind the library, you can read the [introductory notes](#). Visit the [installation page](#) to see how you can download the package. You can browse the [example gallery](#) to see what you can do with seaborn, and then check out the [tutorial](#) and [API reference](#) to find out how.

To see the code or report a bug, please visit the [github repository](#). General support issues are most at home on [stackoverflow](#), where there is a seaborn tag.

Contents

- [Introduction](#)
- [Release notes](#)
- [Installing](#)
- [Example gallery](#)
- [Tutorial](#)
- [API reference](#)

Features

- Relational: [API](#) | [Tutorial](#)
- Categorical: [API](#) | [Tutorial](#)
- Distribution: [API](#) | [Tutorial](#)
- Regression: [API](#) | [Tutorial](#)
- Multiples: [API](#) | [Tutorial](#)
- Style: [API](#) | [Tutorial](#)
- Color: [API](#) | [Tutorial](#)

Viz Options

1. Pandas Visualisation module
2. Matplotlib
3. Seaborn
4. Other options: (Bokeh, Vega, Vincent, Altair, Plotly, ...)

EDA Workflow (Recall...)

1. **Build** a DataFrame from the data (ideally, put all data in this object)
2. **Clean** the DataFrame. It should have the following properties
 1. Each row describes a single object
 2. Each column describes a property of that object
 3. Columns are numeric whenever appropriate
 4. Columns contain atomic properties that cannot be further decomposed
3. Explore **global properties**. Use histograms, scatter plots, and aggregation functions to summarize the data.
4. Explore **group properties**. Use groupby and small multiples to compare subsets of the data.

Types of Visualisations

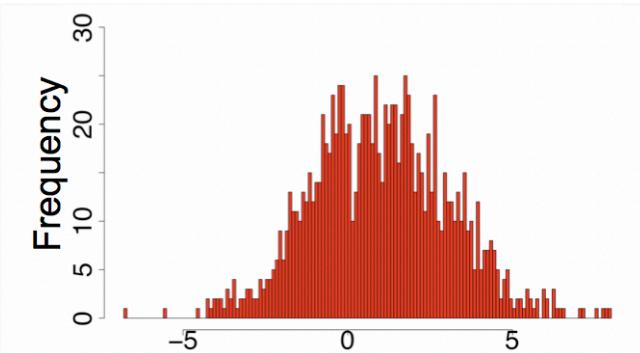
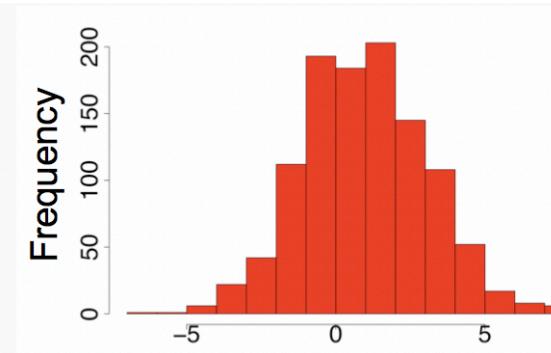
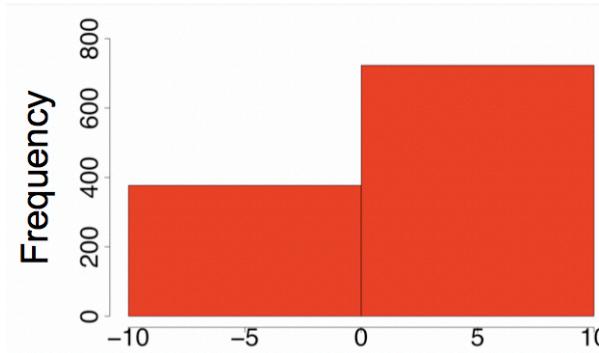
Types of Visualisations

What do you want your visualization to show about your data?

- **Distribution:** how a variable or variables in the dataset distribute over a range of possible values.
- **Relationship:** how the values of multiple variables in the dataset relate
- **Composition:** how the dataset breaks down into subgroups
- **Comparison:** how trends in multiple variable or datasets compare

Histograms

A **histogram** is a way to visualize how 1-dimensional data is distributed across certain values.

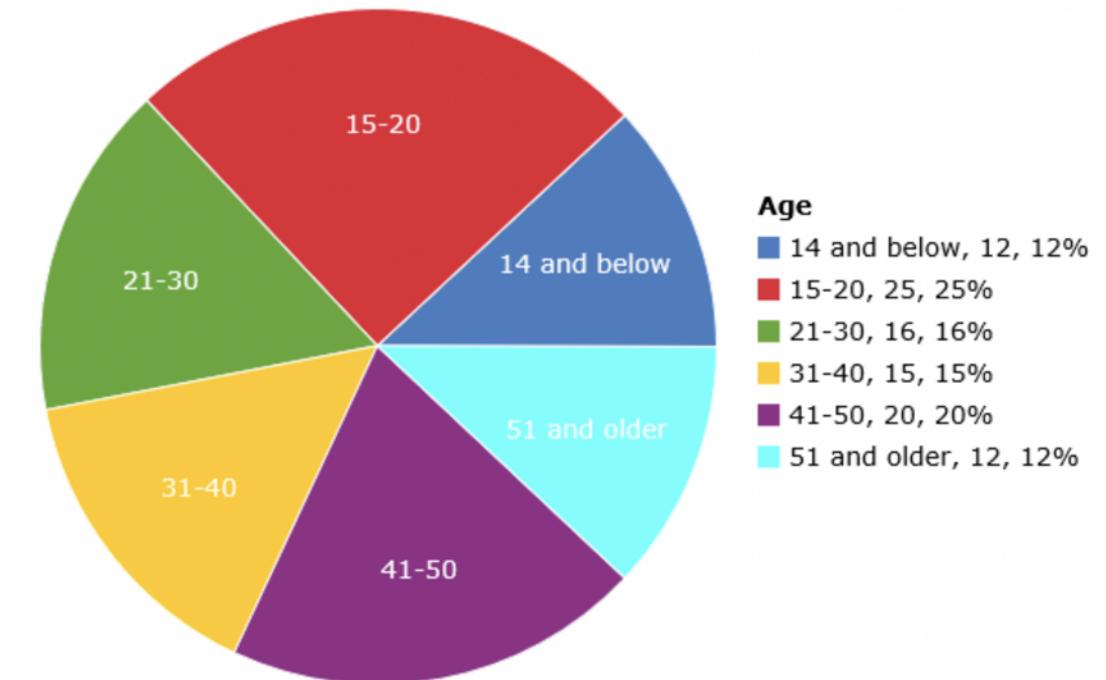


Note: Trends in histograms are sensitive to number of bins.

Pie Charts for Categorical Variables

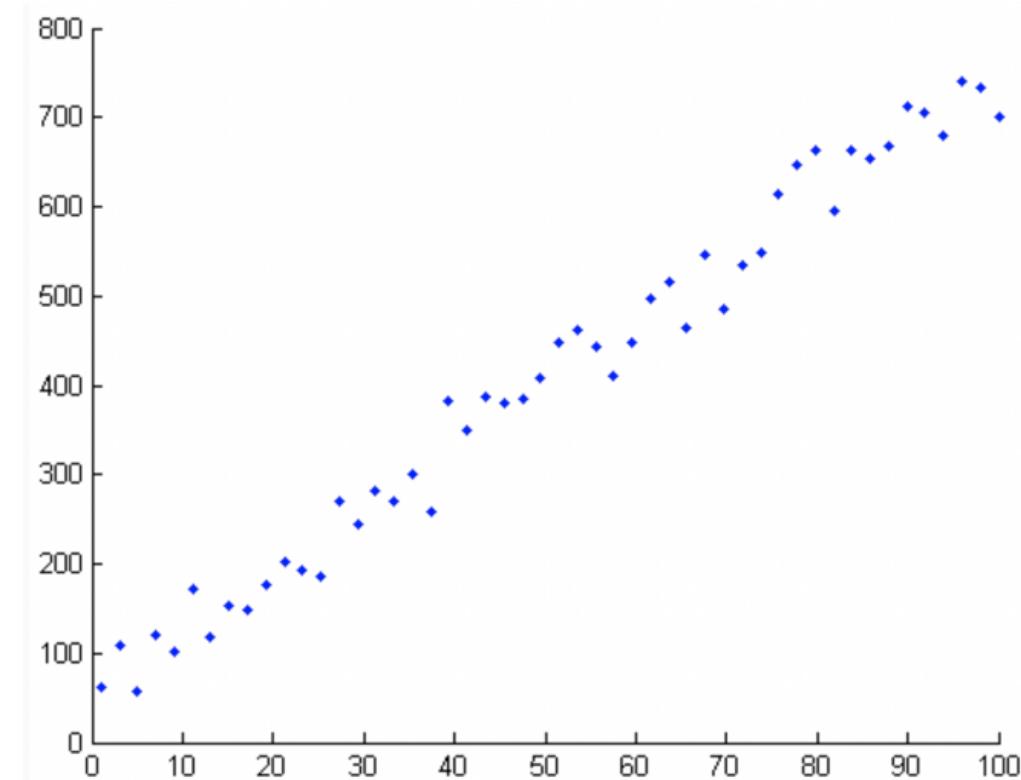
A **pie chart** is a way to visualize the static composition (aka, distribution) of a variable (or single group).

Pie charts are often frowned upon (and bar charts are used instead). Why?



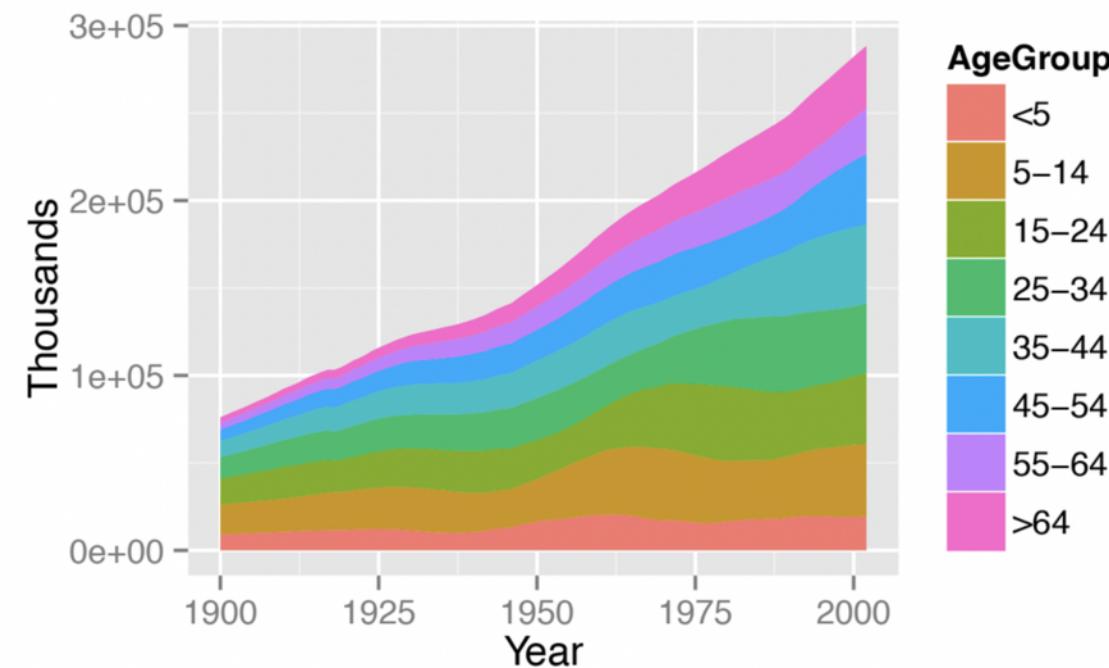
Scatter Plots to Visualise Relationships

A **scatter plot** is a way to visualize the relationship between two different attributes of multi-dimensional data.



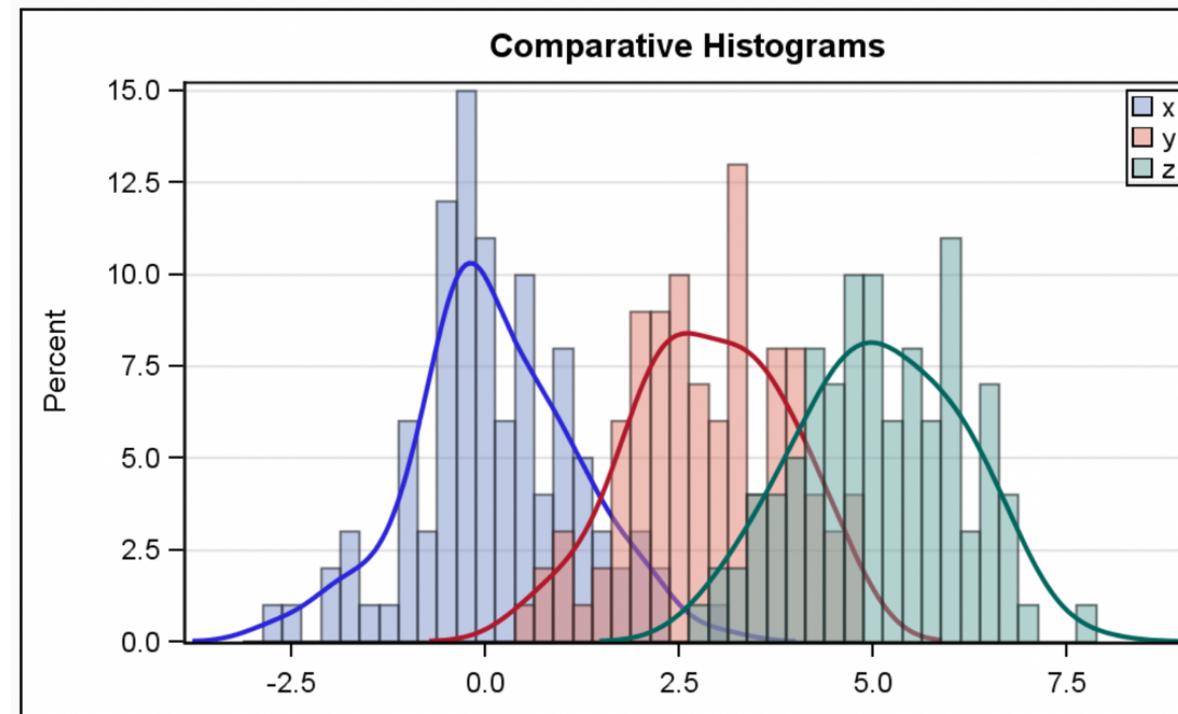
Stacked area graph to show trend over time

A **stacked area graph** is a way to visualize the composition of a group as it changes over time (or some other quantitative variable). This shows the relationship of a categorical variable (AgeGroup) to a quantitative variable (year).



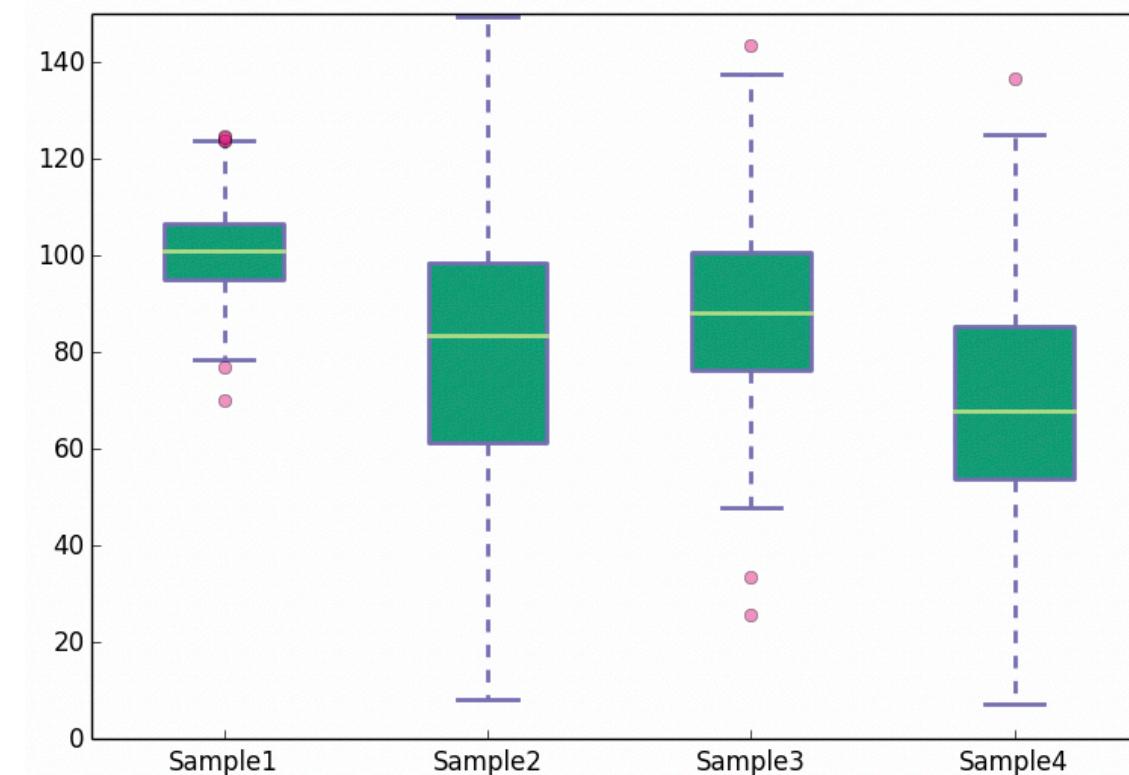
Multiple Histograms

Plotting **multiple histograms** (and **kernel density estimates** of the distribution, here) on the same axes is a way to visualize how different variables compare (or how a variable differs over specific groups).



Boxplots

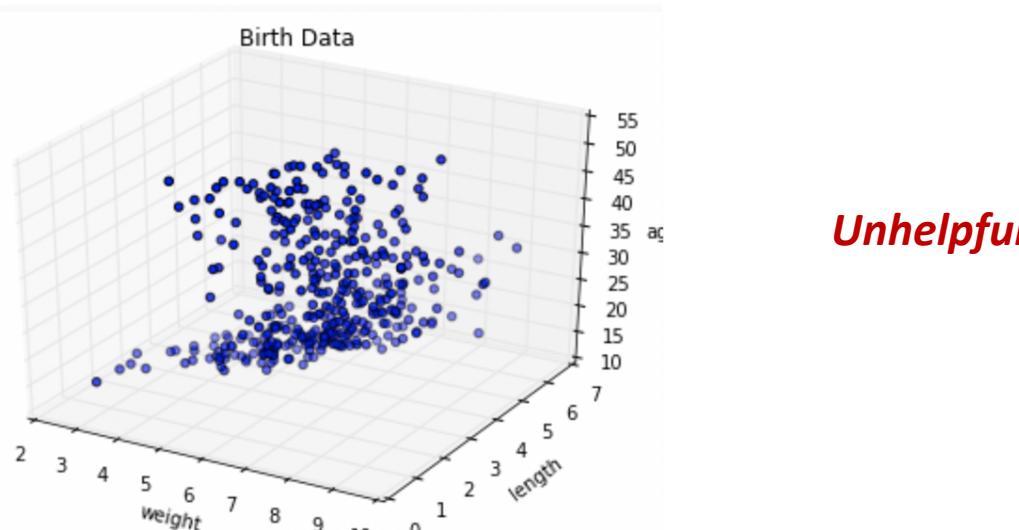
A **boxplot** is a simplified visualization to compare a quantitative variable across groups. It highlights the range, quartiles, median and any outliers present in a data set.



Not Everything is Possible!

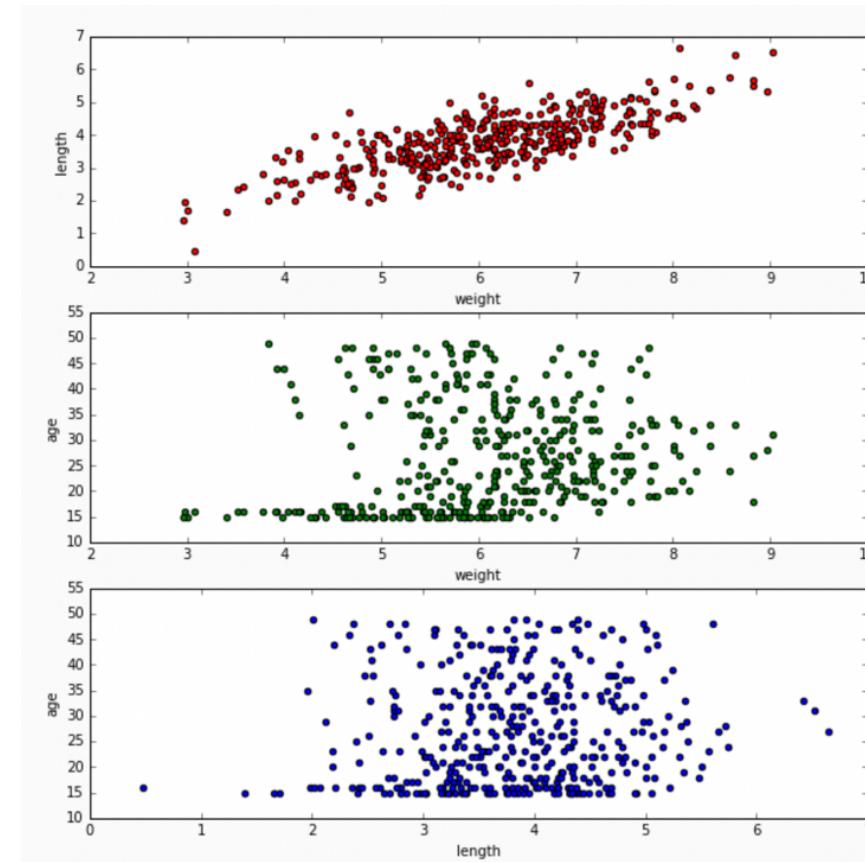
Often your dataset seem too complex to visualize:

- Data is too high dimensional (how do you plot 100 variables on the same set of axes?)
- Some variables are categorical (how do you plot values like Cat or No?)



Reducing Complexity

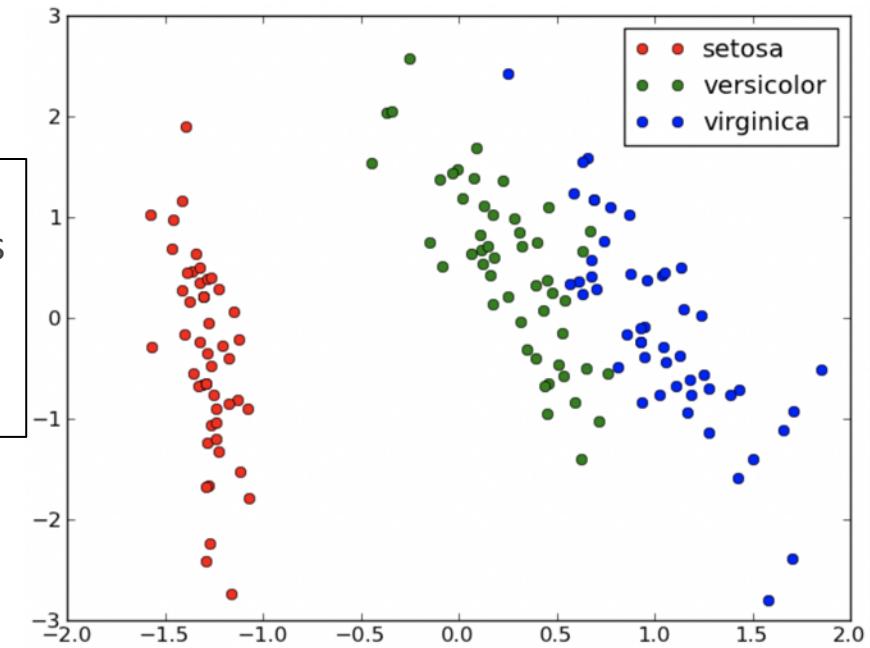
Relationships may be easier to spot by producing multiple plots of lower dimensionality.



Reducing Complexity

For 3D data, color coding a categorical attribute can be “effective”

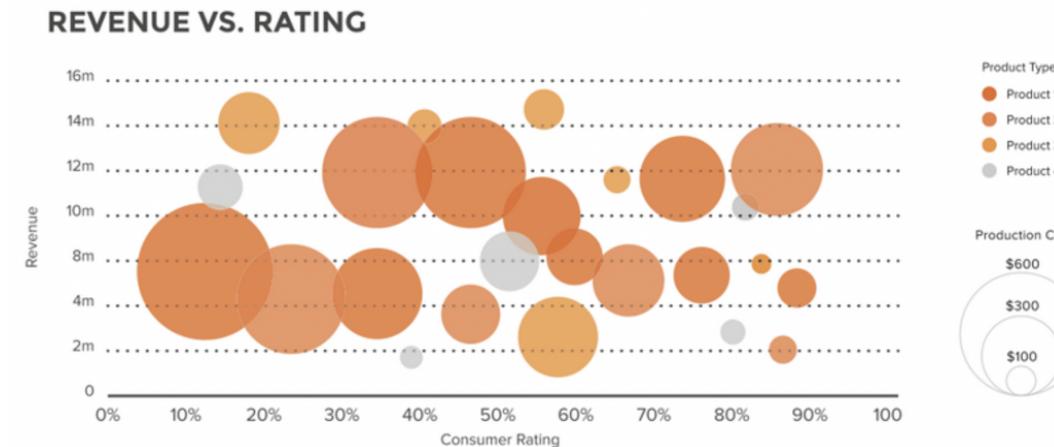
This visualizes a set of Iris measurements. The variables are petal length, sepal length, Iris type (setosa, versicolor, virginica).



Except when it's not effective.
What could be a better choice?

3D can work

For 3D data, a quantitative attribute can be encoded by size in a bubble chart.



The above visualizes a set of consumer products. The variables are revenue, consumer rating, product type and product cost.

Break - Design Exercise (Time: 10 min)

Q: How Do You Feel about doing science?

Interest	Before	After
Excited (E)	19	38
kind of E	25	30
Ok	40	14
Not great	5	6
Bored	11	12

Instructions

1. **What do you want to do:** Analyse data or Communicate an insight
2. Sketch a visualisation (pen and paper is fine)
3. Take a photo and submit on Assignments in Brightspace under *Visualisation: Design Exercise* (.jpg, .jpeg, .png)
4. Submission deadline tonight by 2330
5. Discussion of some of your submissions follows in Lecture 06.
6. Exercise is **not** graded

After the pilot program,

68%

of kids expressed interest towards science,
compared to 44% going into the program.

Break



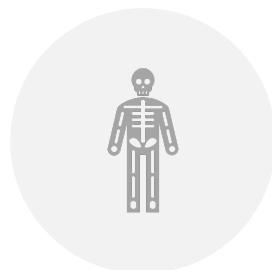
WATER



WALK



COFFEE OR TEA

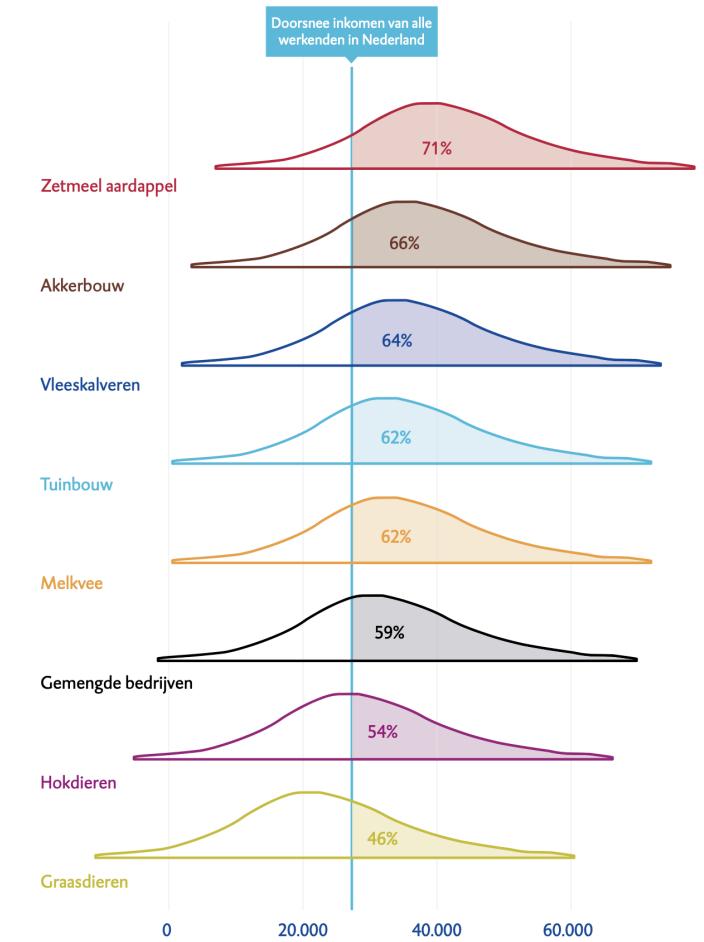


MAKE FRIENDS

Effective Visualisation

“The greatest value of a picture is when it forces us to notice what we never expected to see”

John Tukey



[Source]

Treasury Quarterly Net Marketable Borrowing
"Net Cash"
Fiscal Quarter

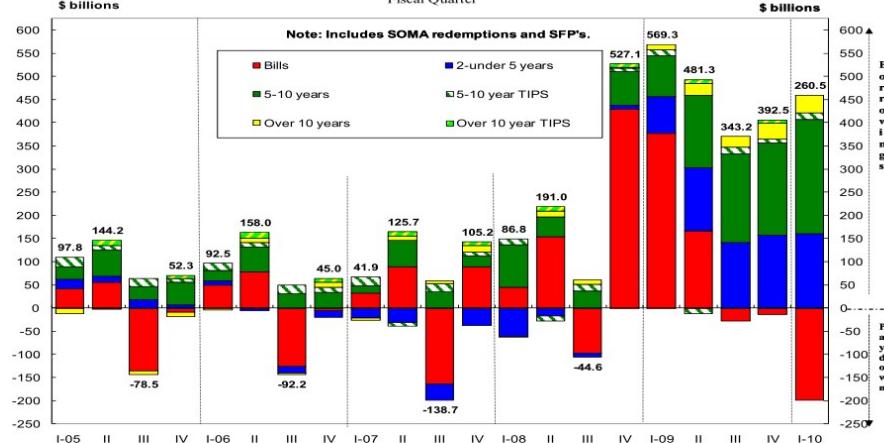
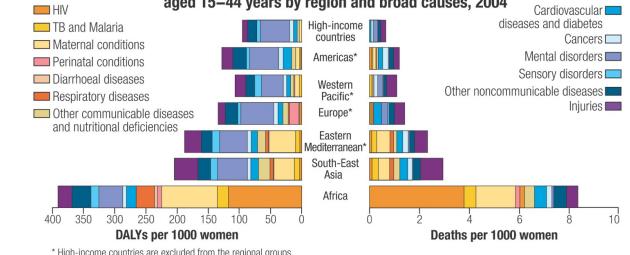


Figure 1 Mortality and disease burden (DALYs) in women aged 15–44 years by region and broad causes, 2004



* High-income countries are excluded from the regional groups.
Source: World Health Organization.

Not Effective...

Sources: US Treasury and WHO reports

FY 2012 Total Liabilities (Composition)

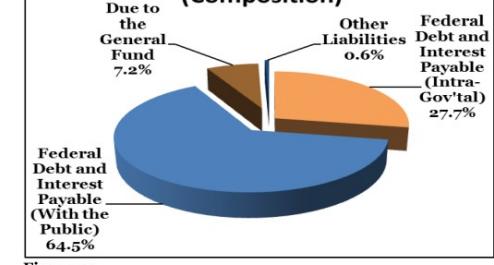
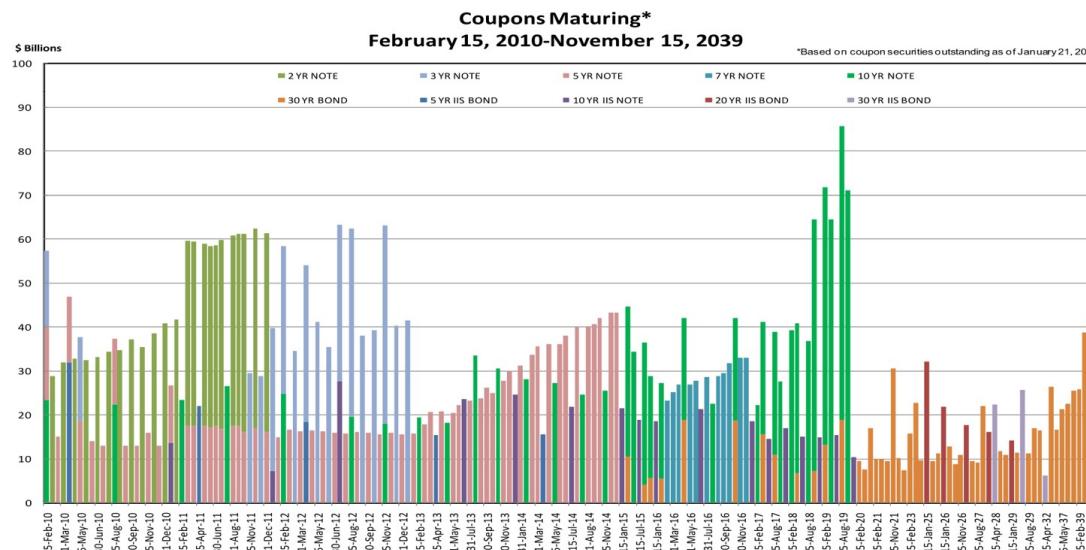


Figure 10



Cryptosporidium Prevalence

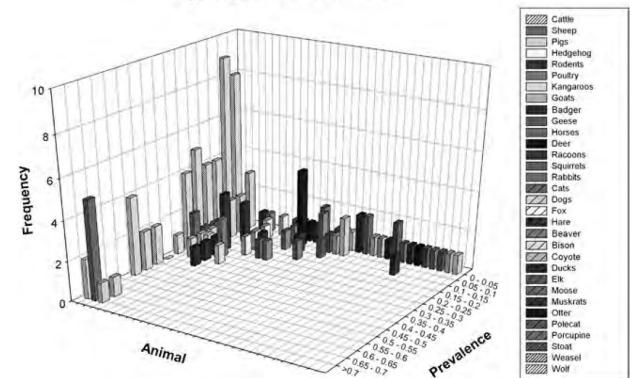


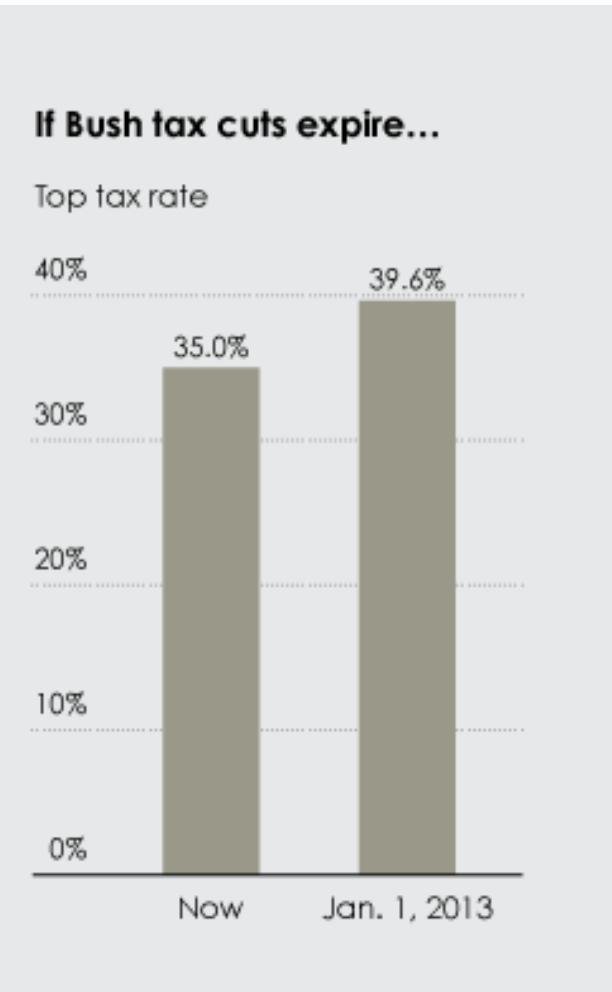
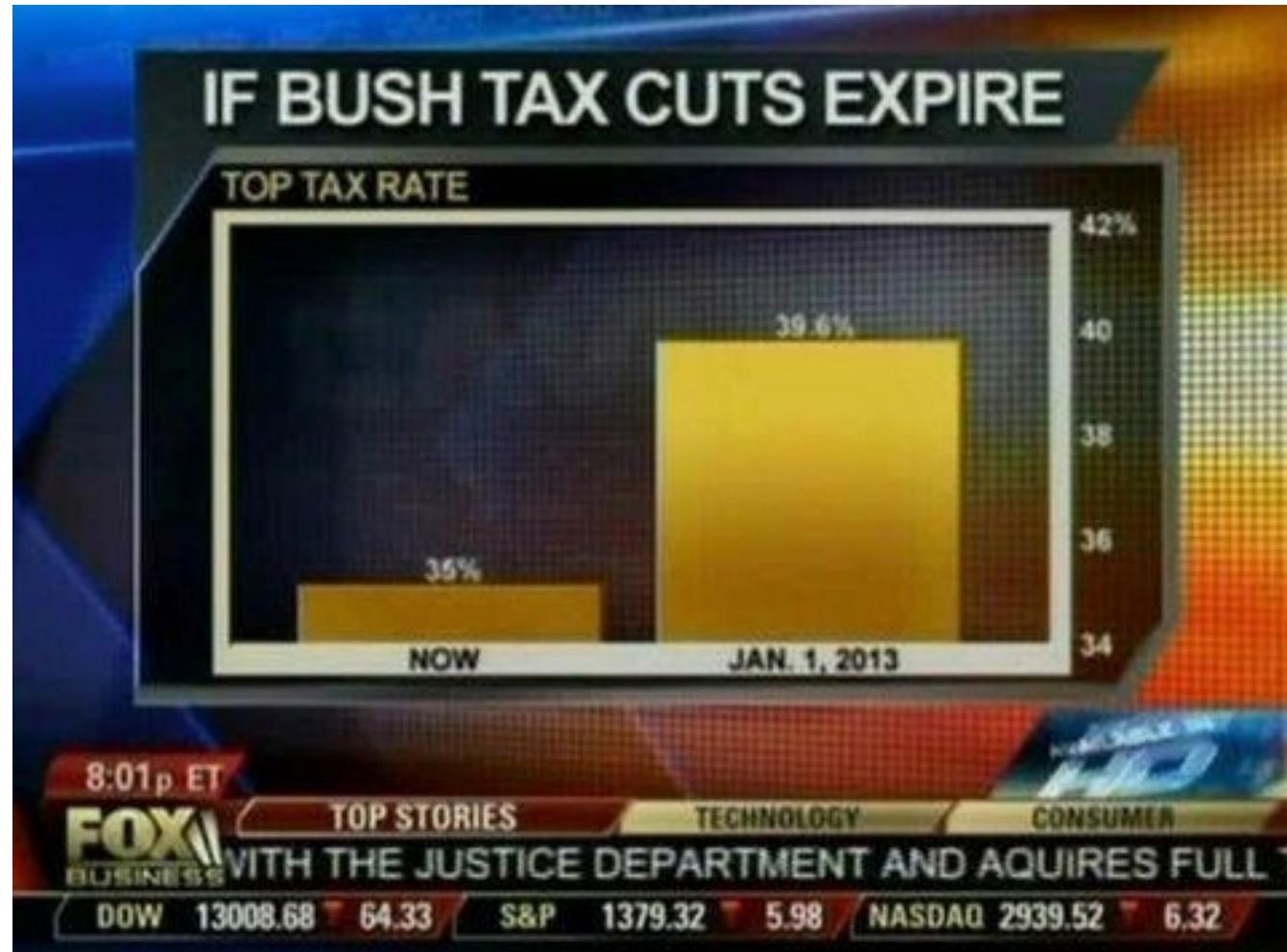
Figure 5.2 Mean prevalence rates of *Cryptosporidium* oocysts by animal species.

Effective EDA Visualisation

1. Have graphical integrity
2. Keep it simple
3. Use the right display
4. Use colour strategically
5. *Tell a story with data*

1. Graphical Integrity

Scale Distortions

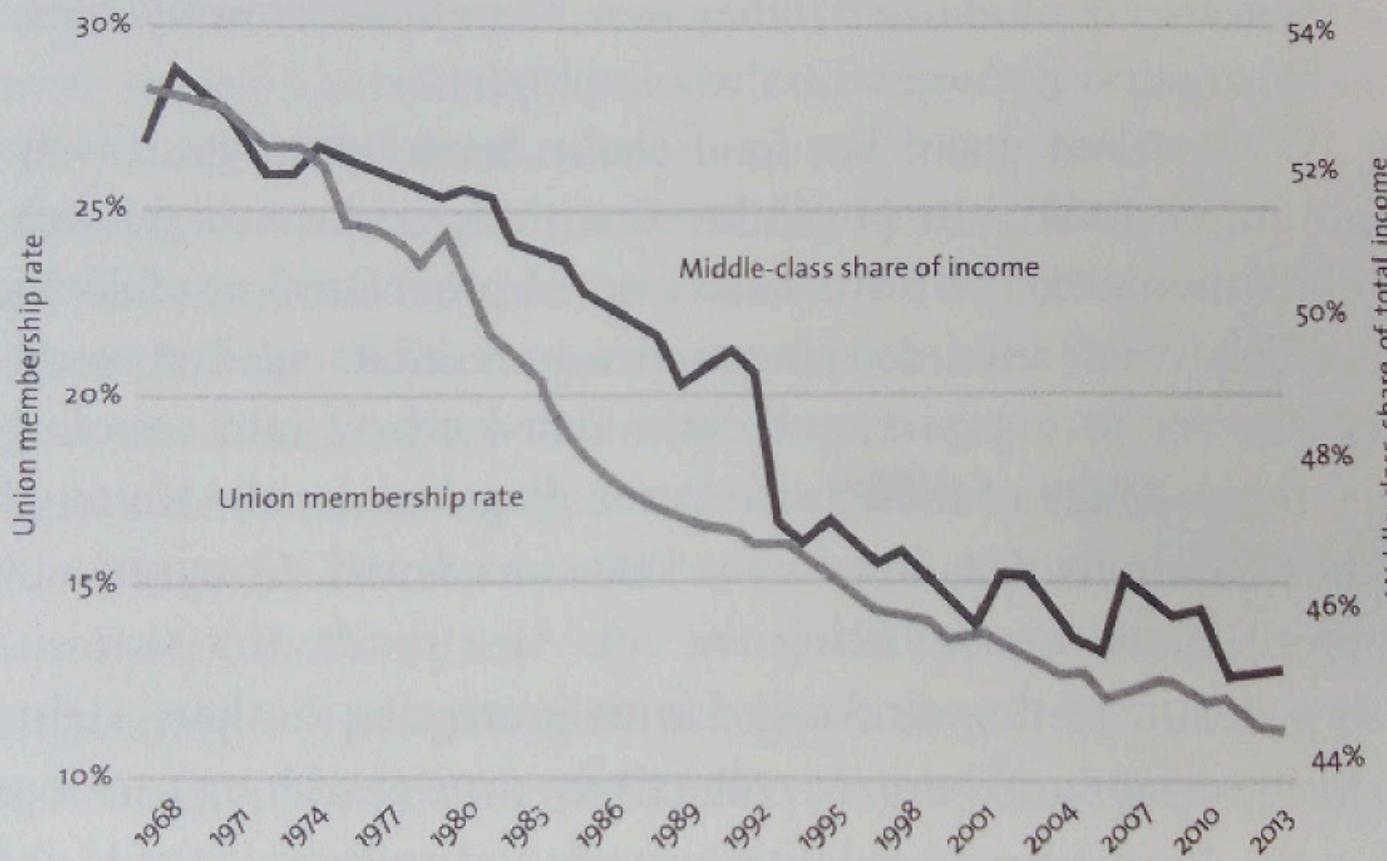


Scale Distortions

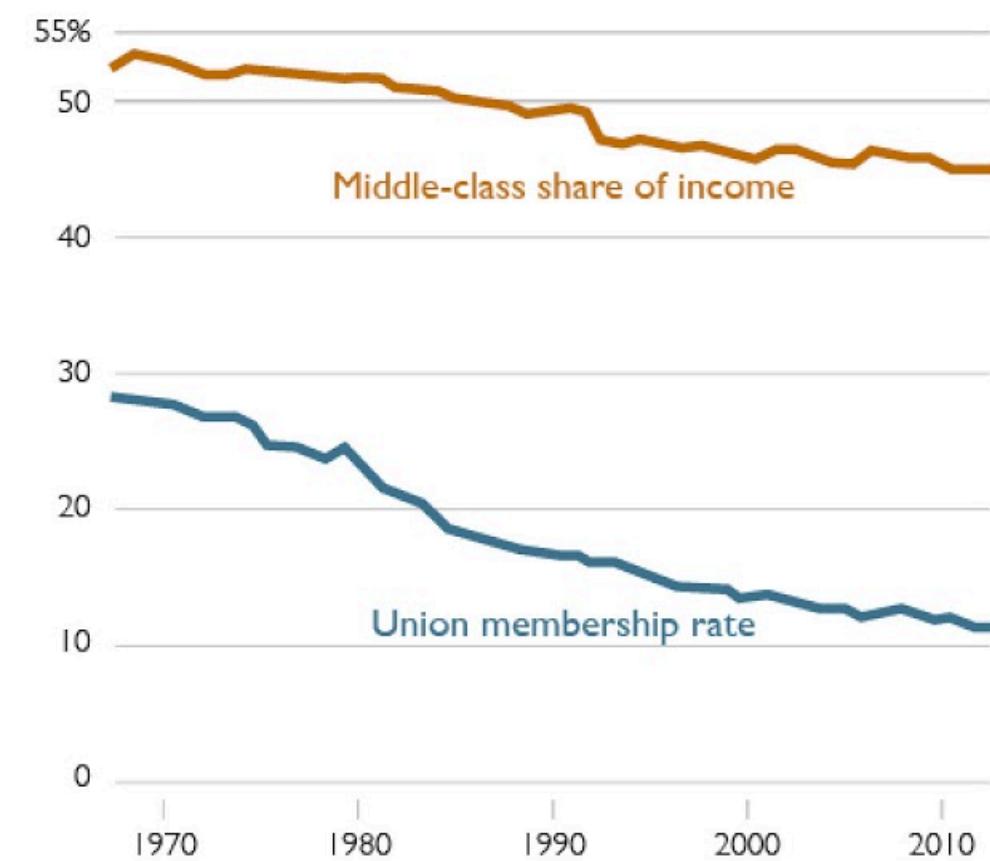


“Double the axes, double the mischief”

FIGURE 7. AS UNION MEMBERSHIP DECLINES, THE SHARE OF INCOME GOING TO THE MIDDLE CLASS SHRINKS



NEW VERSION



Include Uncertainty

Think about Perceptions

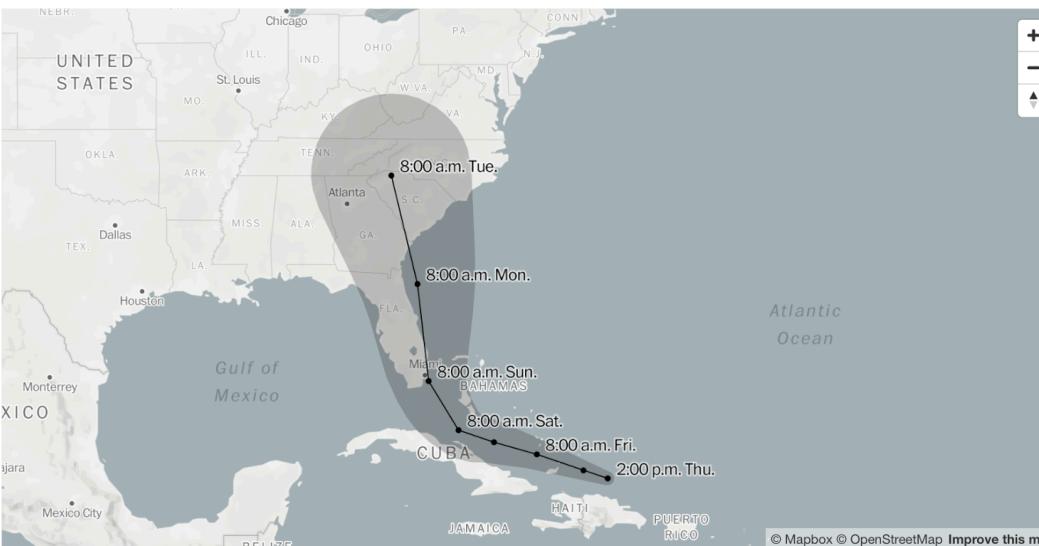
National

What's in the path of Hurricane Irma

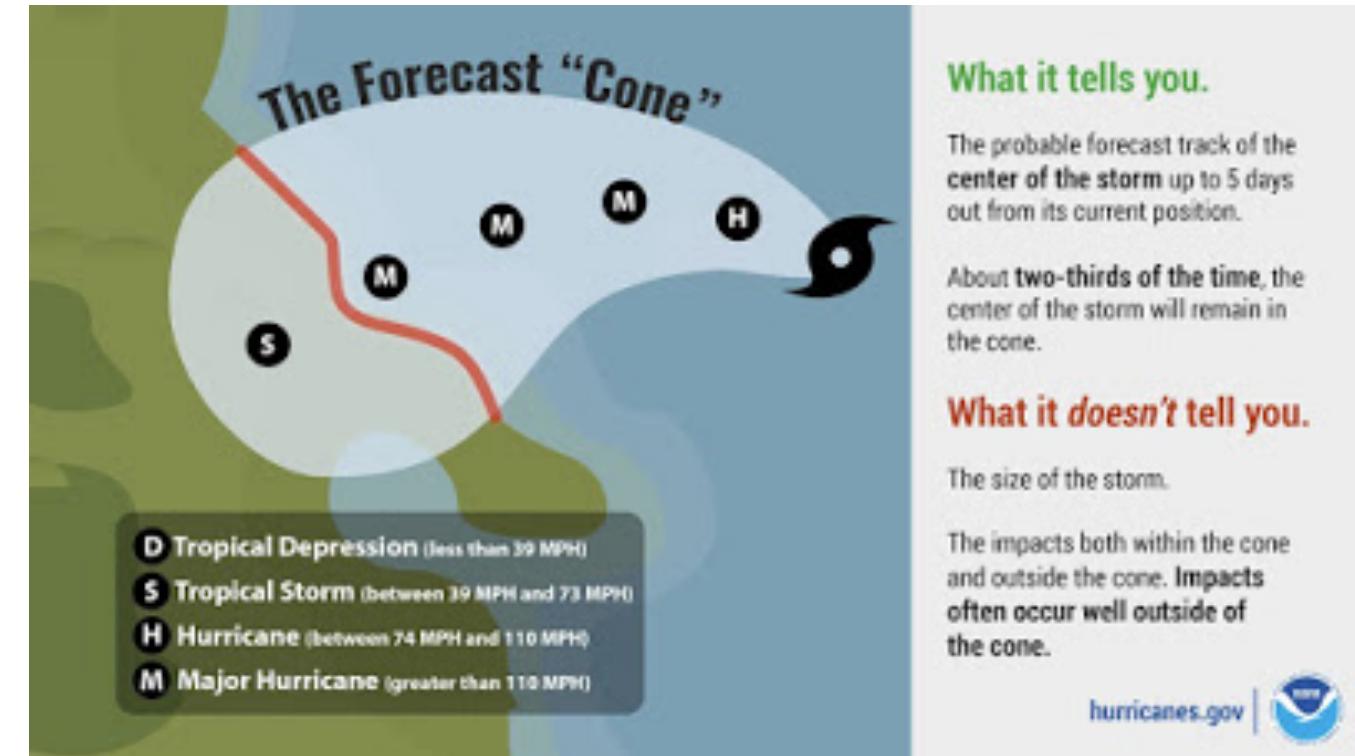
The storm is big, fast and speeding through the Caribbean toward Florida

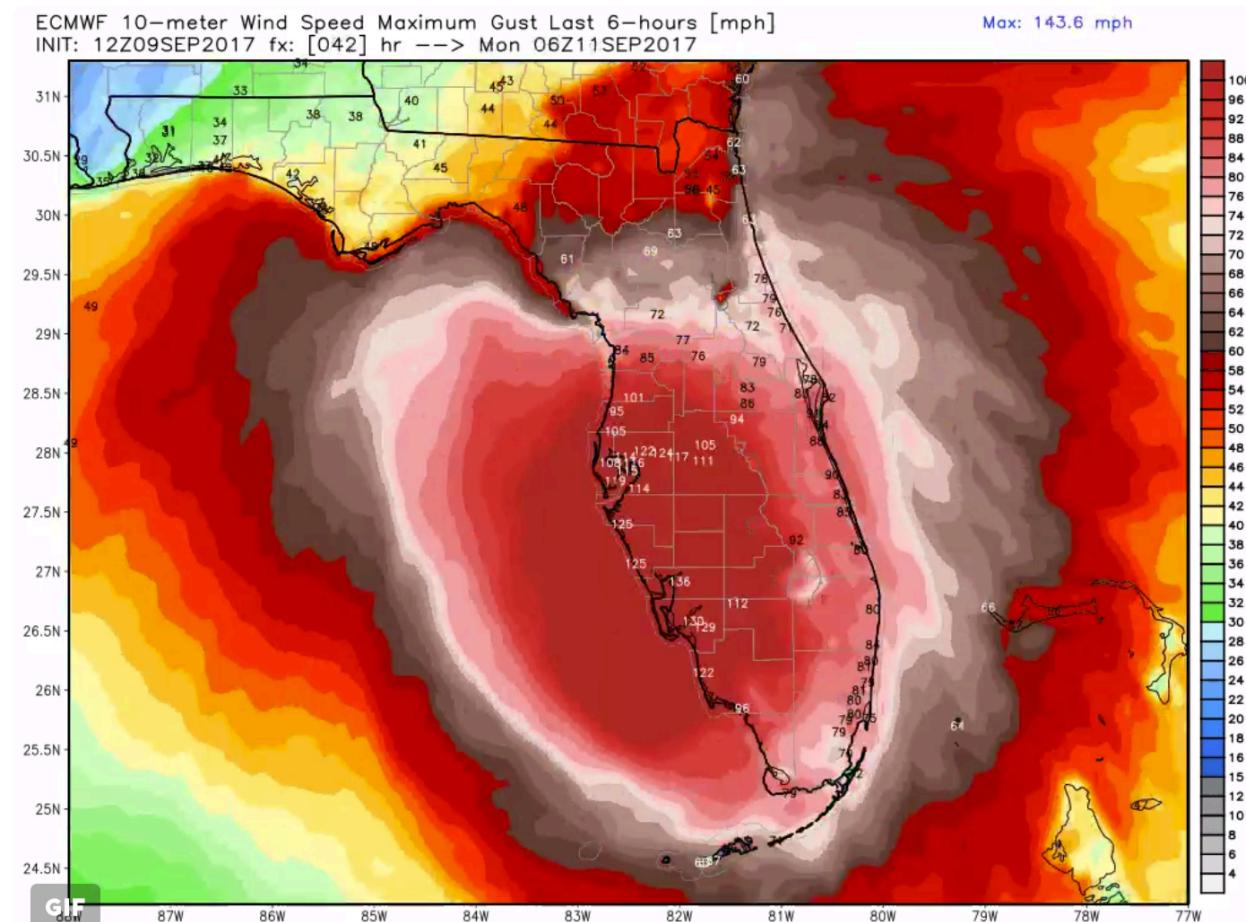
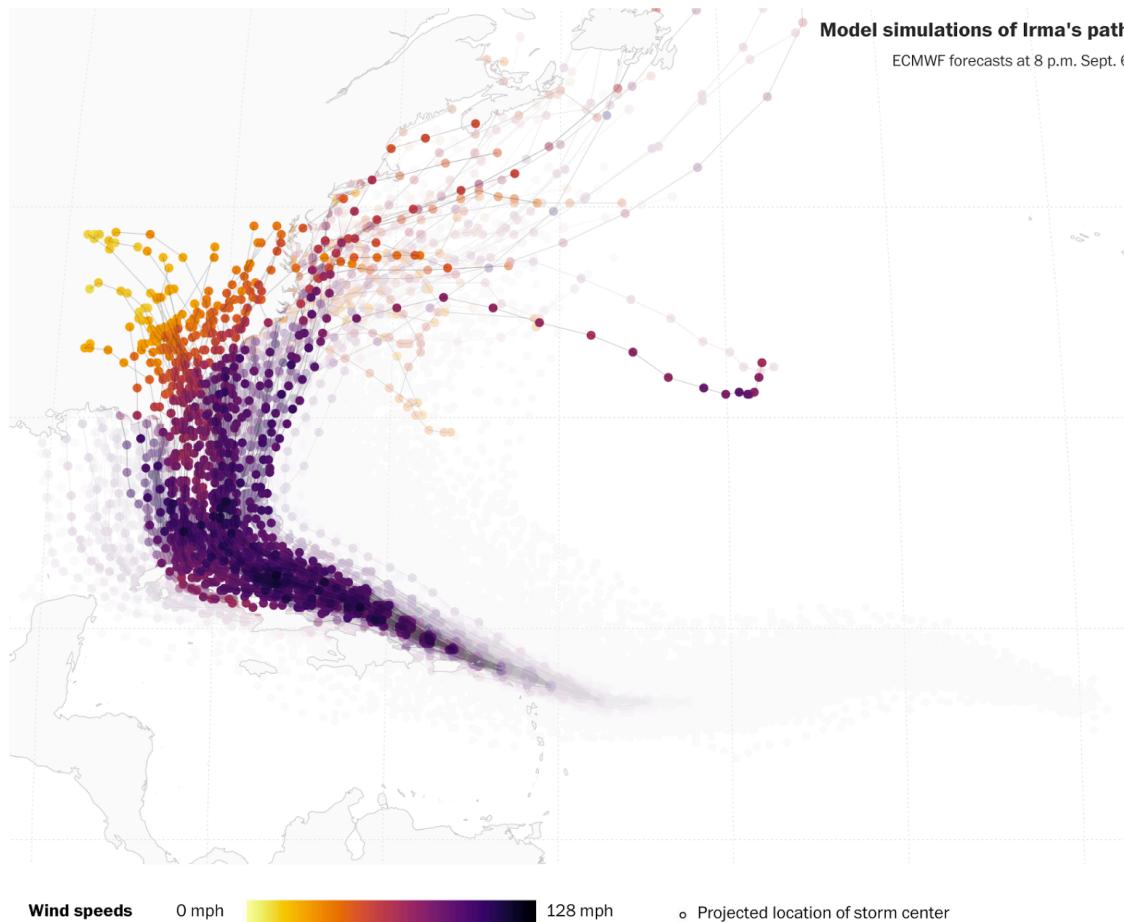
By Bonnie Berkowitz, Laris Karklis, Reuben Fischer-Baum, John Muyskens, Gabriel Florit and Denise Lu

Updated Sept. 7 3:15 p.m.



The future path of the center of Hurricane Irma could be anywhere within this cone.



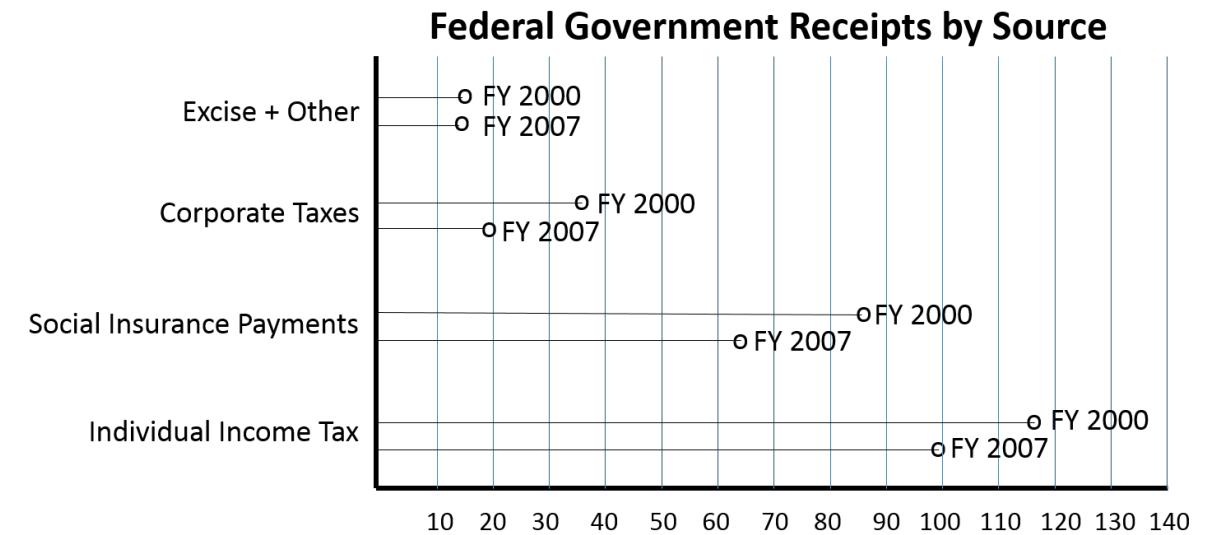
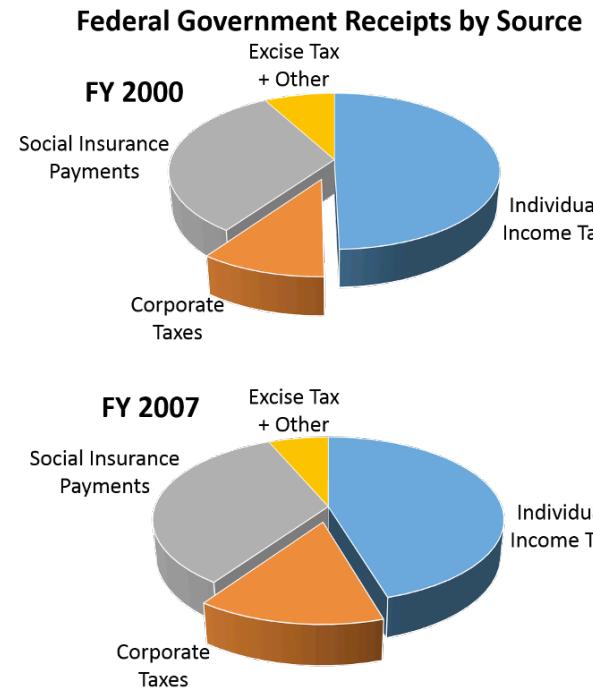


2. Keep it simple

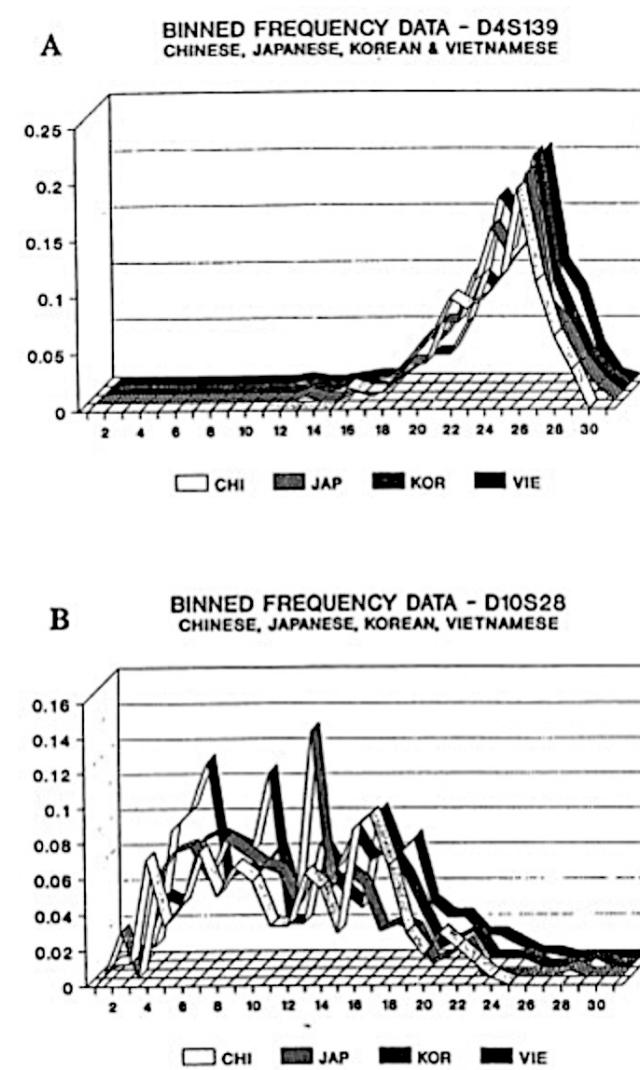
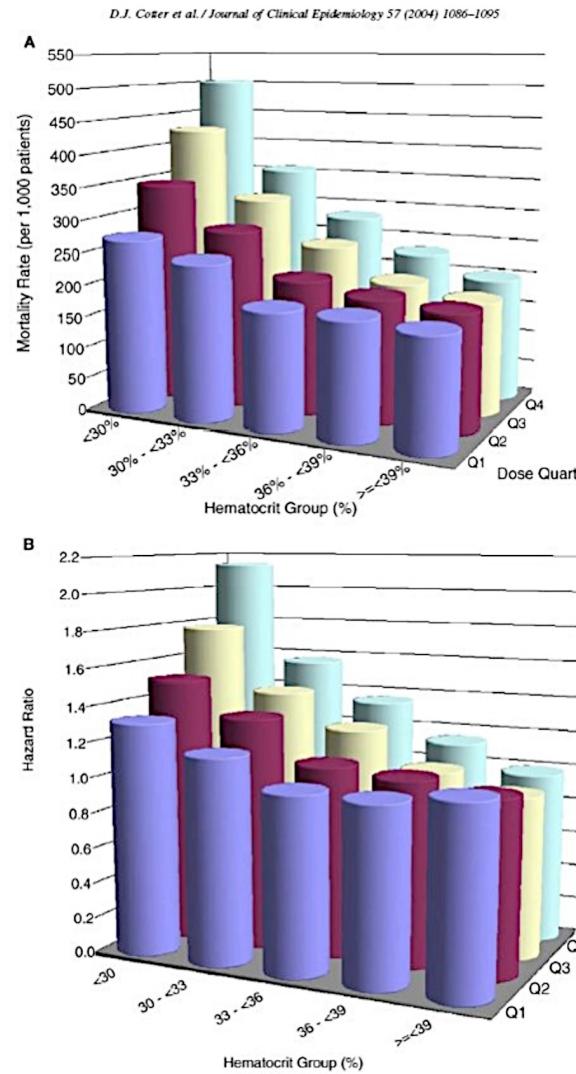
Maximise Data-Ink Ratio – show GIF!

Remove
to improve
(the **data-ink** ratio)

The use of Pie Charts is generally discouraged



Exclude unneeded dimensions



Exclude unneeded dimensions

Much easier to make comparisons

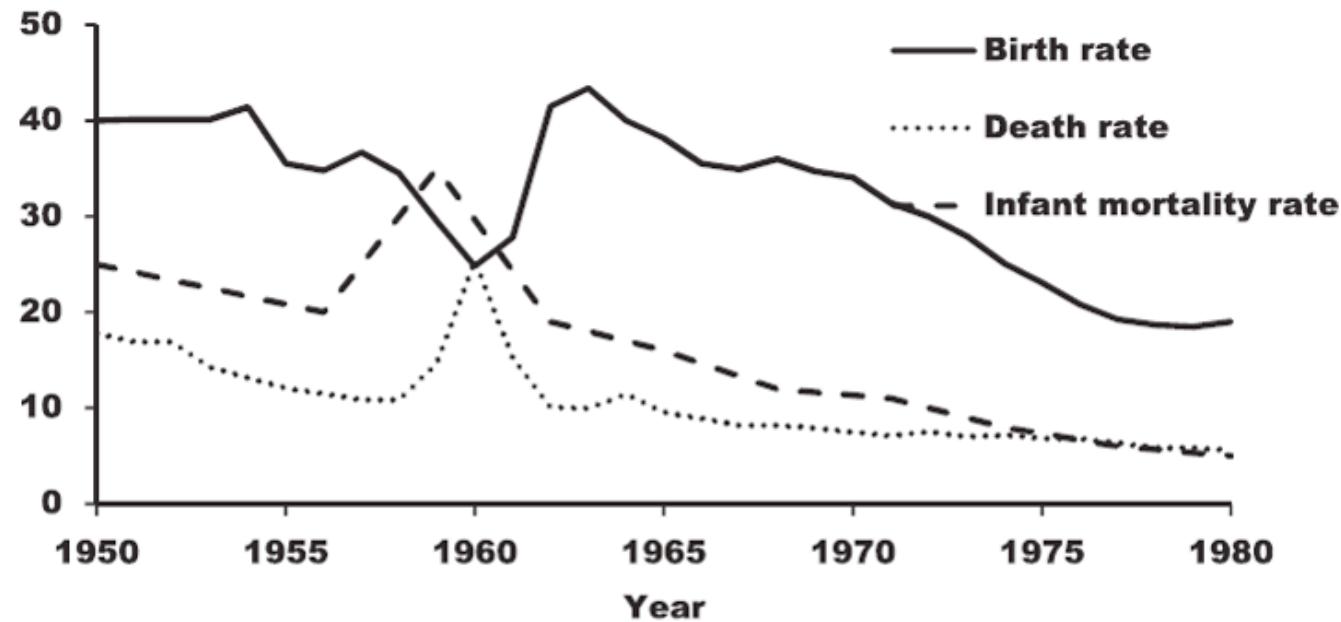
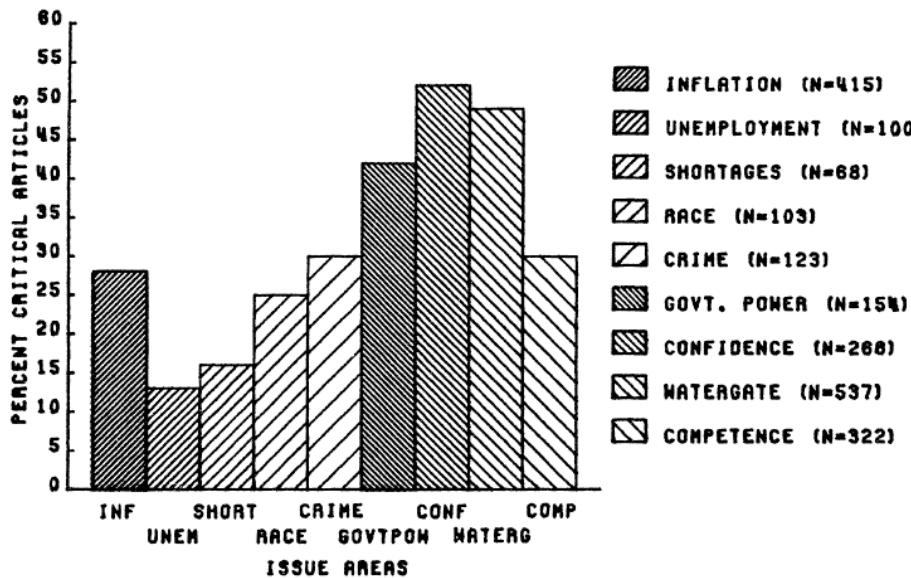


Figure 1. Fertility reduction and excess death rate and infant mortality (per thousand) during the Chinese Famine of 1959-61.

Sources: computed from the 1982 Population Census of China and the 1988 Two-Per-Thousand National Survey of Fertility and Contraception.

Omit chart junk



Source: Center for Political Studies Media Content Analysis Study, 1974; available through the University of Michigan, ICPSR. Not to be cited without full bibliographical reference to the present article.

- Unnecessary bar graphs
- Pointless and annoying cross-hatching labeled with incomplete abbreviation - Moiré
- Difficult to go back and forth from the legend to the bar graph
- All uppercase letters are hard to read

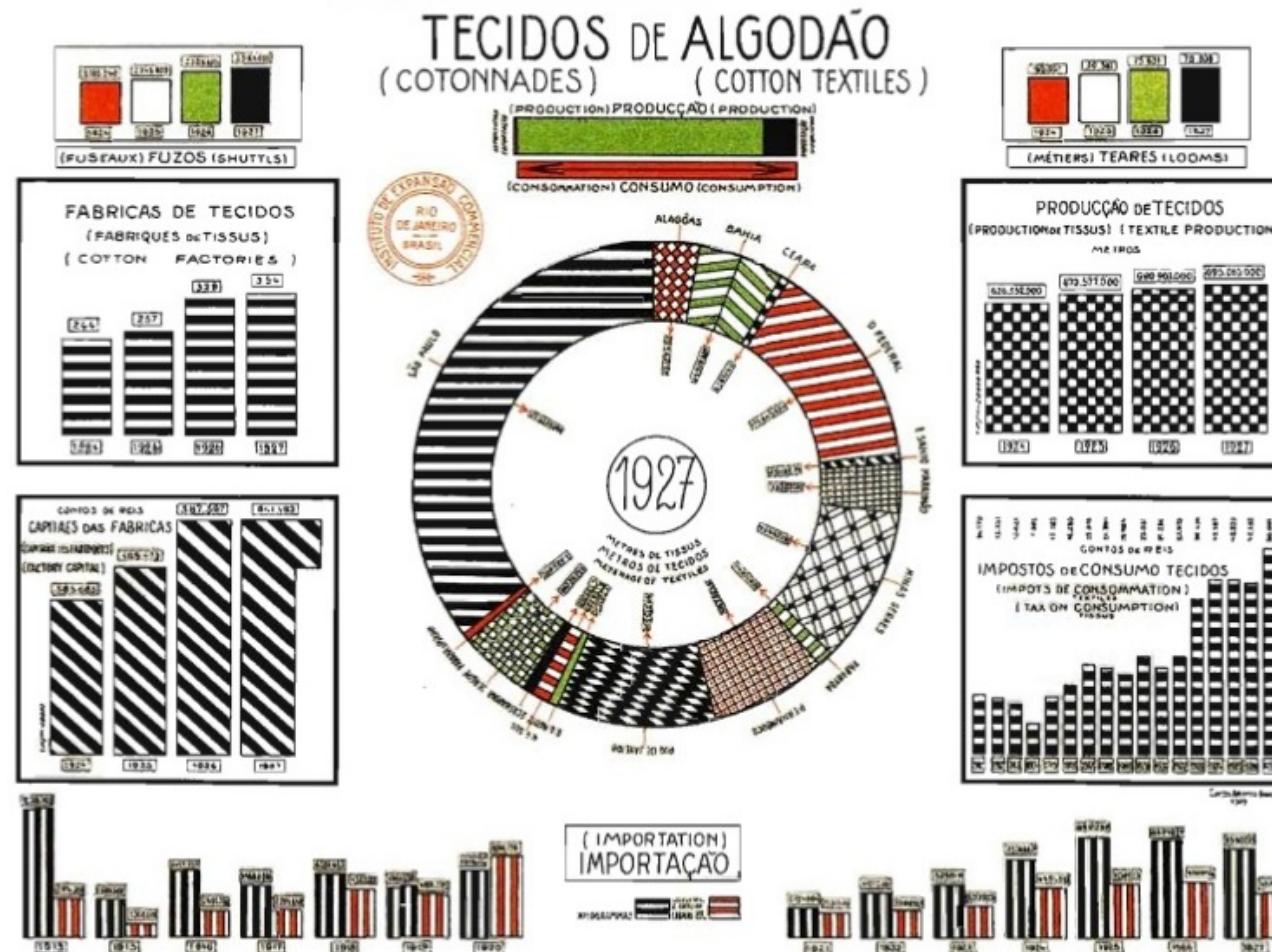
Omit chart junk

No reason to connect these counts with lines



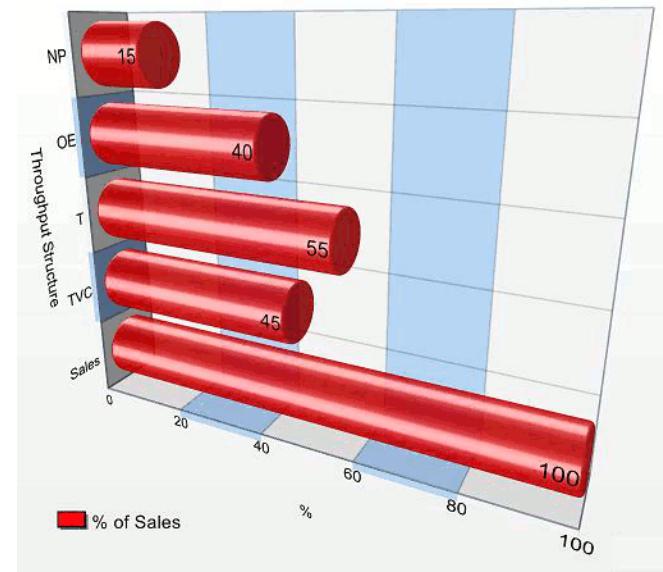
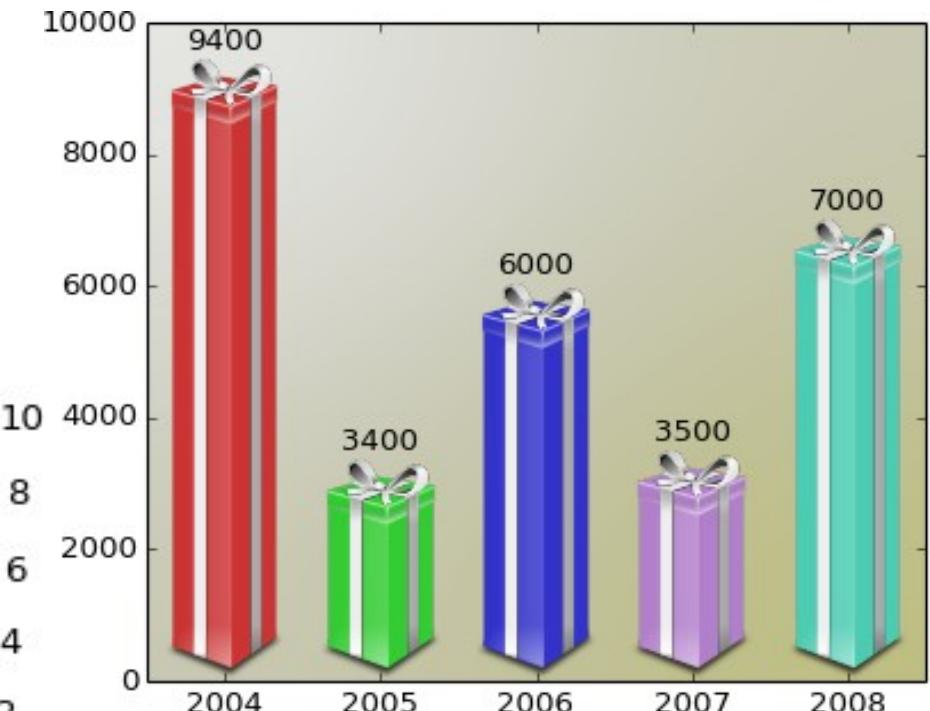
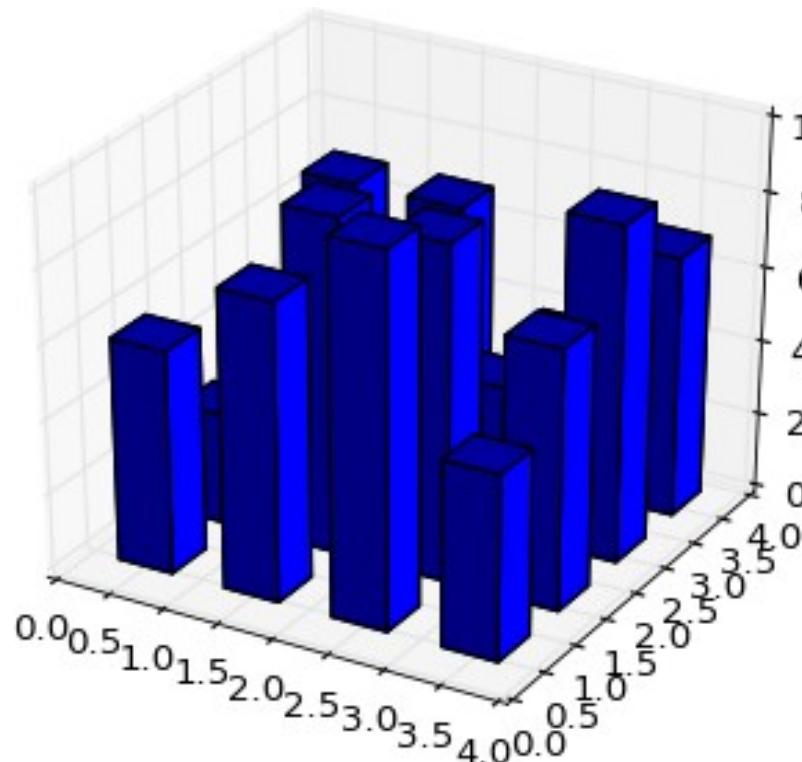
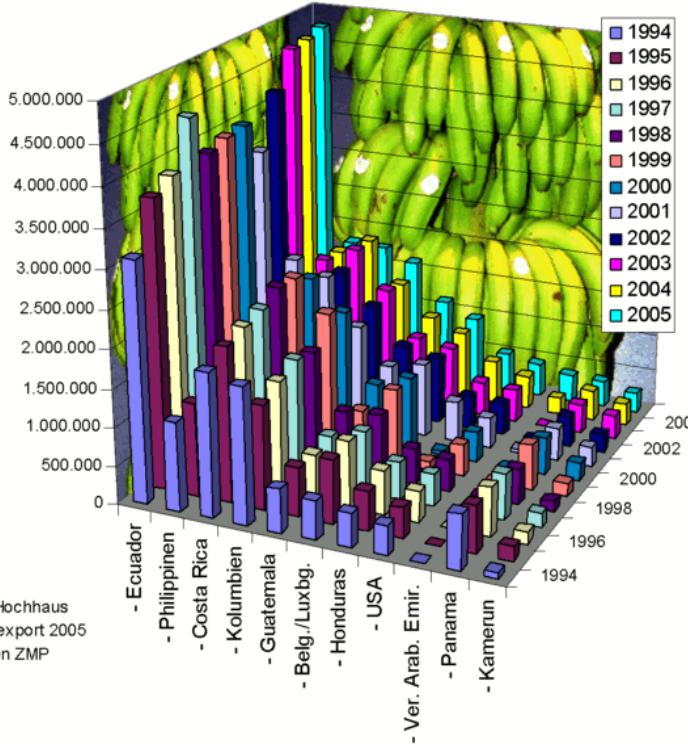
Omit chart junk

Moiré vibration – visual noise, distracting



Don't!

Export von Bananen in Tonnen von 1994-2005



3. Use the right display

Deviation

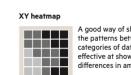
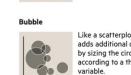
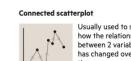
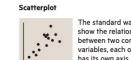
Emphasise variations (+/-) from a fixed reference point. It can also mean a point is zero but can also be a target or a long-term average. Can also be used to show sentiment (positive/negative).

Example FT uses
Trade surplus/deficit, climate change

**Correlation**

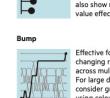
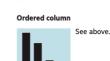
Show the relationship between two or more variables. If one variable goes up, tell them otherwise, many readers will assume the relationships they show to be causal (i.e. one causes the other).

Example FT uses
Inflation and unemployment, income and life expectancy

**Ranking**

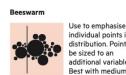
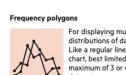
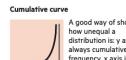
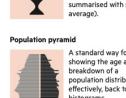
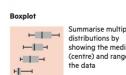
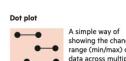
Use where an item's position in an ordered list is more important than its absolute or raw value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results

**Distribution**

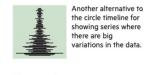
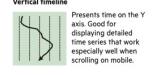
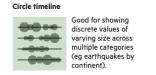
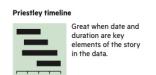
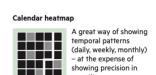
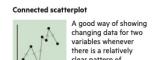
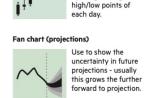
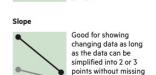
Show values in a dataset and how often they occur (frequency or skewness). This can be a good way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population (geographical), distribution, revealing

**Change over Time**

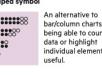
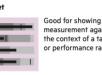
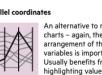
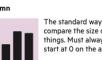
Give emphasis to changing trends. Those that are short (eg day-to-day) moments or extended periods traversing decades or centuries. Choosing the correct time period is important to provide suitable context for the reader.

Example FT uses
Share price movements, economic time series, sectoral changes in a market

**Magnitude**

Show size comparisons. These can be relative (part-to-whole) or absolute (size). If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Fiscal budgets, company structures, national election results

**Part-to-whole**

Show how a single entity can be broken down into its components. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

**Spatial**

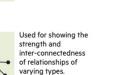
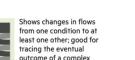
Aside from locator maps only used when the location of data matters or conditions. These might be logical sequences or geographical locations.

Example FT uses
Population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

**Flow**

Show the reader volumes or intensity of movement or flow between entities or conditions. These might be logical sequences or geographical locations.

Example FT uses
Population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results



Visual vocabulary

Designing with data

There are so many ways to visualise data - how do we know which one to pick? Use the categories across the top to decide which data relationship is most important in your story, then look at the different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations.

FT graphic: Alan Smith; Chris Campbell; Ian Bell; Liz Pearce; Graham Farrel; Billy Ehrenberg-Shannon; Paul McCallum; Martin Stabe
Inspired by: Graphic Continuum by Jon Scheidt and Steven Rizzo



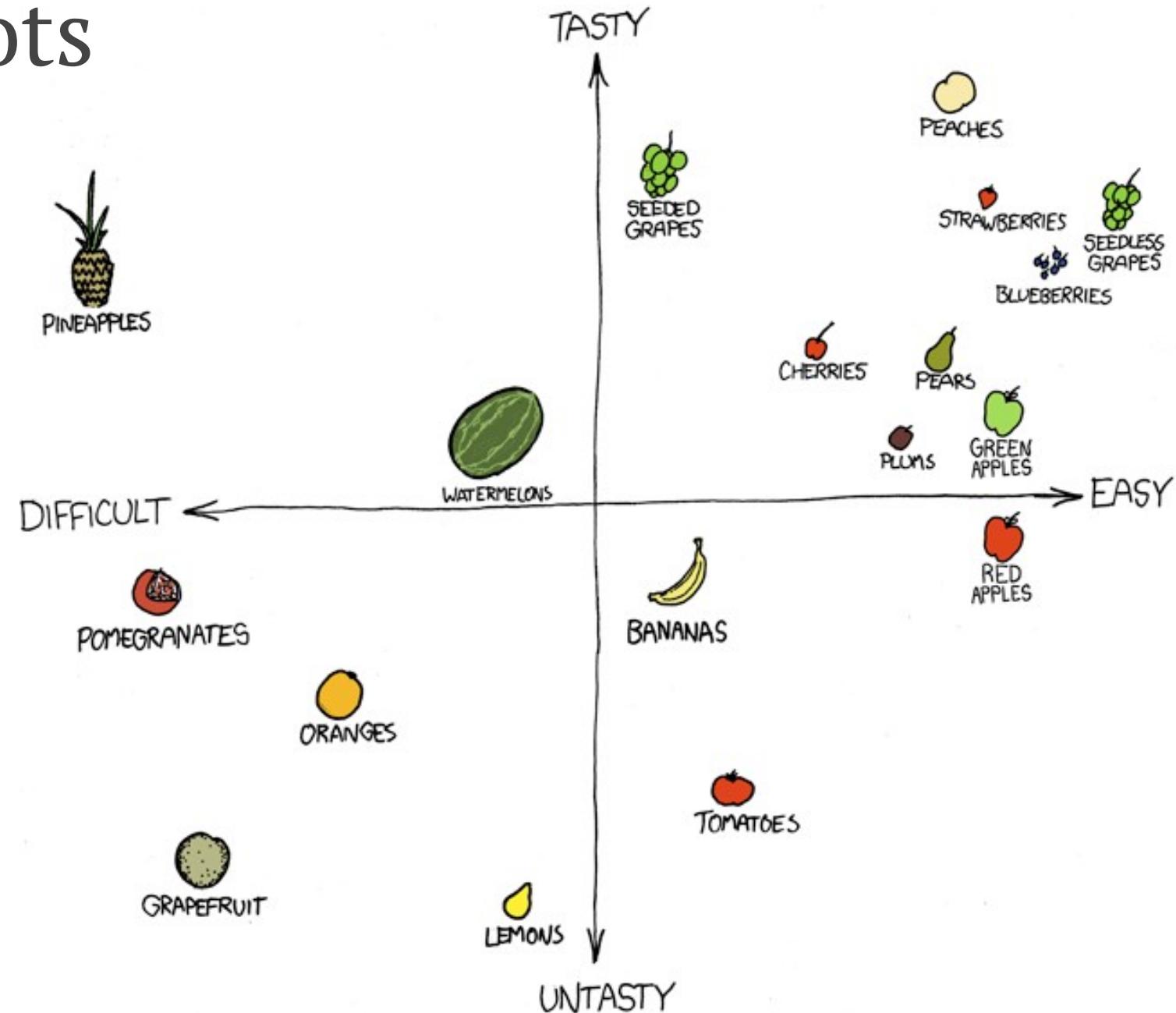
ft.com/vocabulary

FT

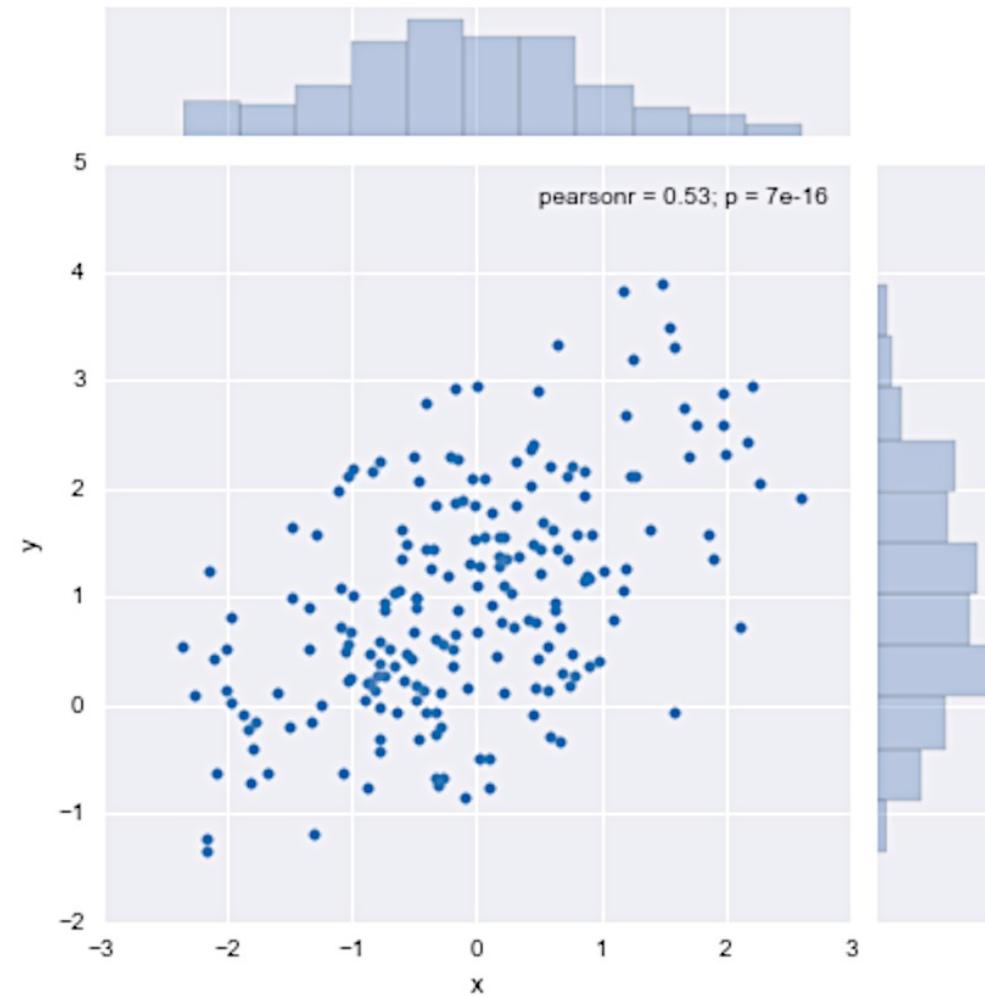
© Financial Times 2014-2015
This work is licensed under a Creative Commons
Attribution-ShareAlike 4.0 International License.

Correlations

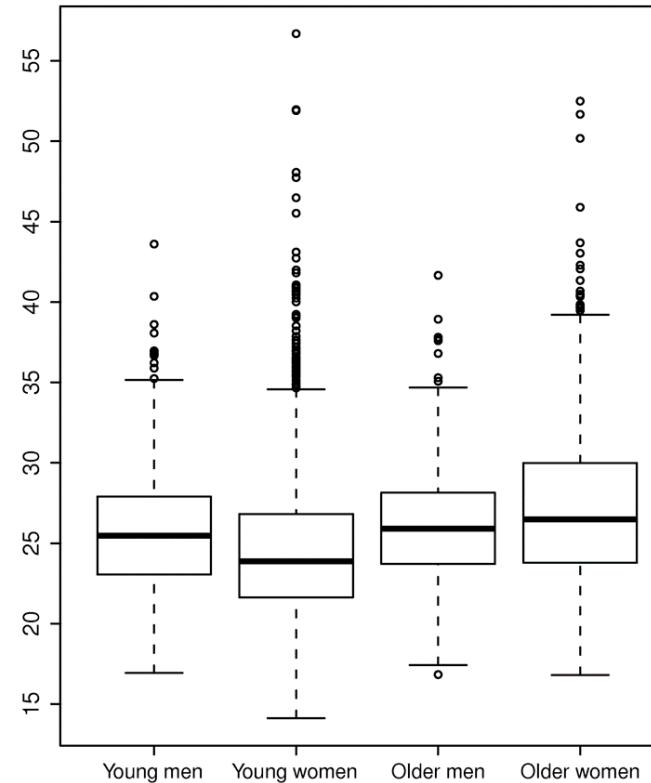
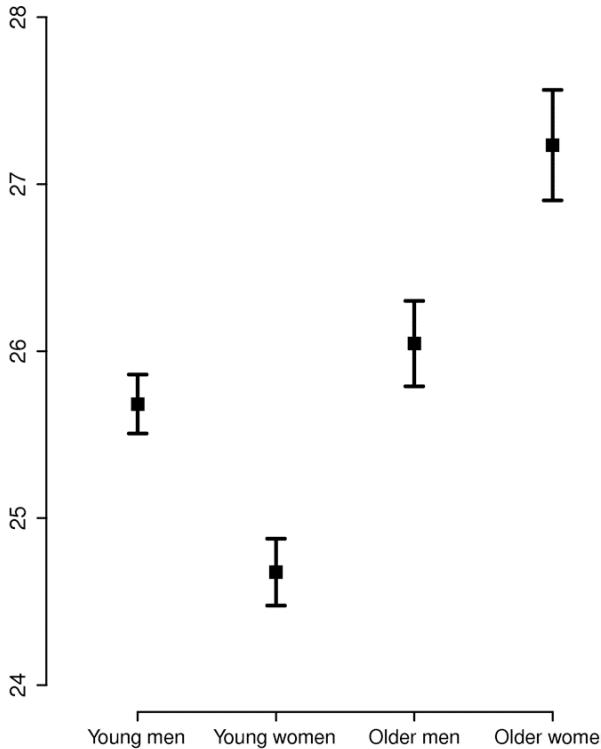
Scatterplots



Make efficient use of space

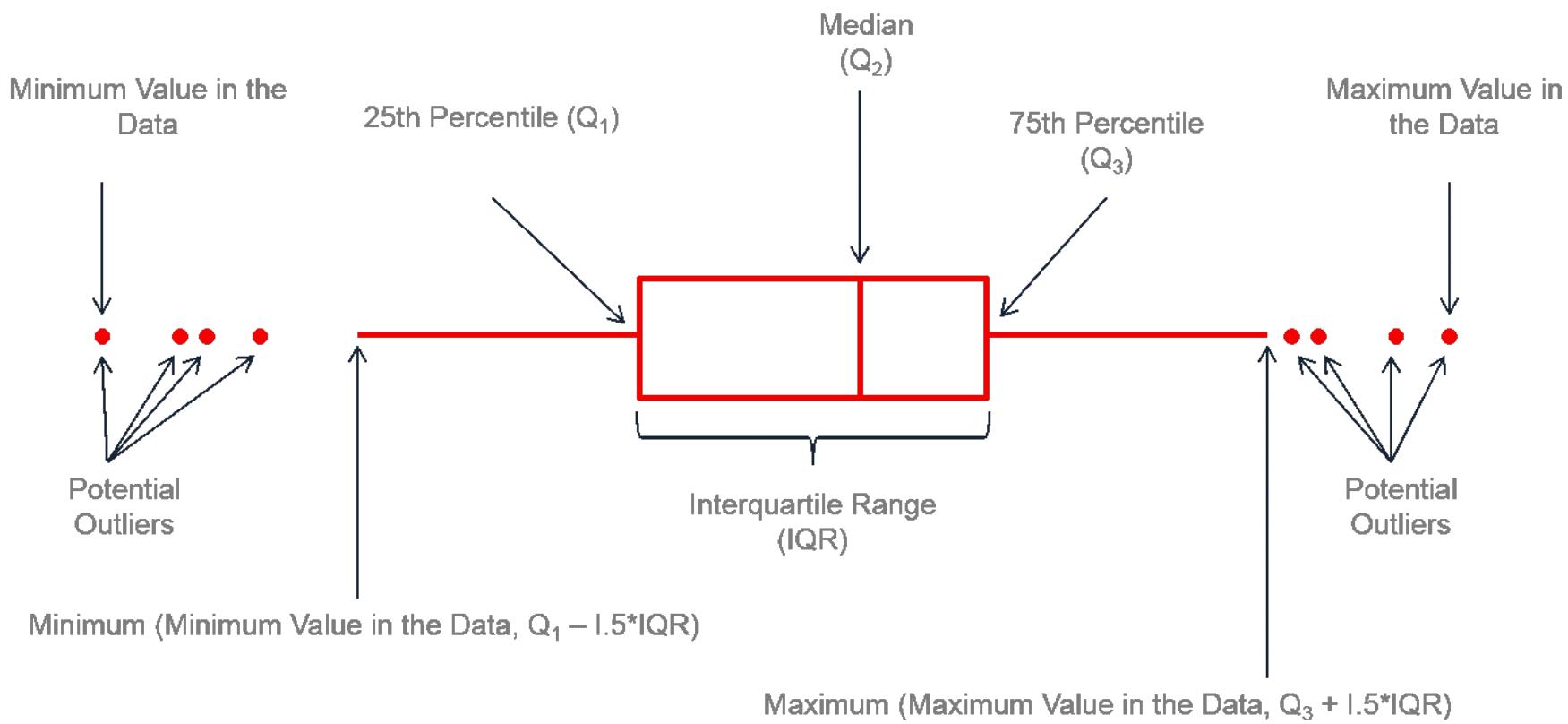


Make efficient use of space



- Error bars for BMI (Body Mass Index) measurements in four categories.
- Left: This is easy to interpret, but the viewer cannot see that the data is quite skewed.

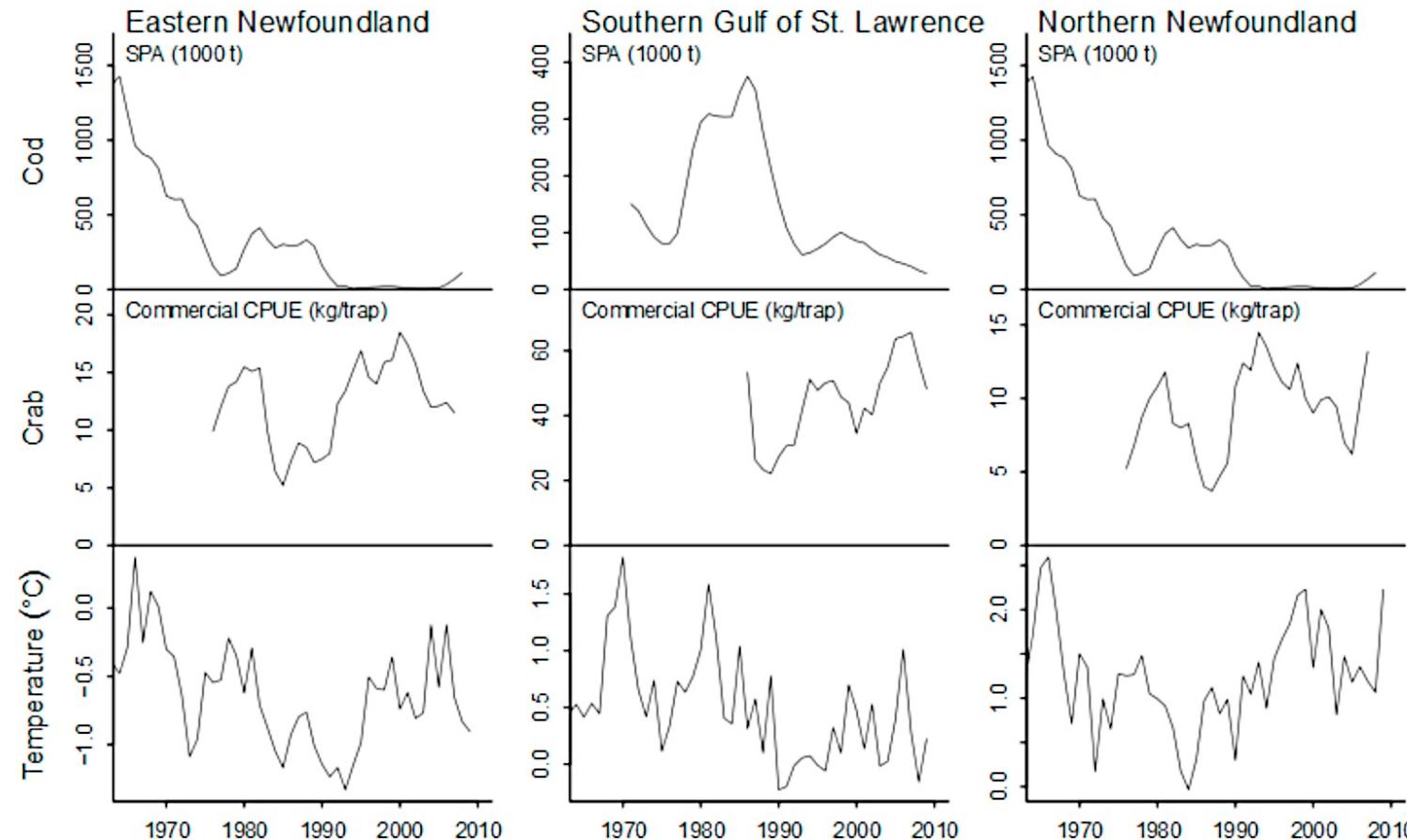
Make efficient use of space



Bar charts are not appropriate for indicating means \pm SEs: they add ink without conveying any additional information

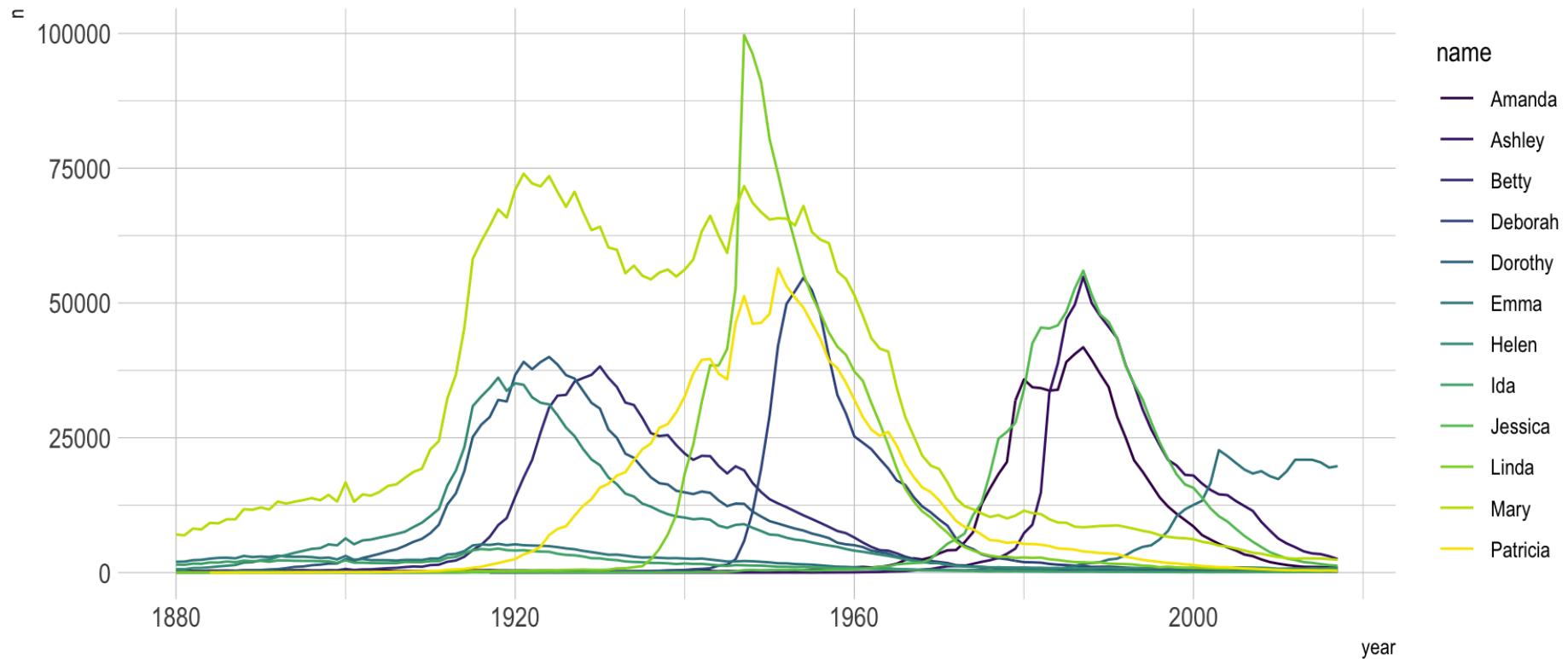
Facilitate Comparison

[Y] axes are different

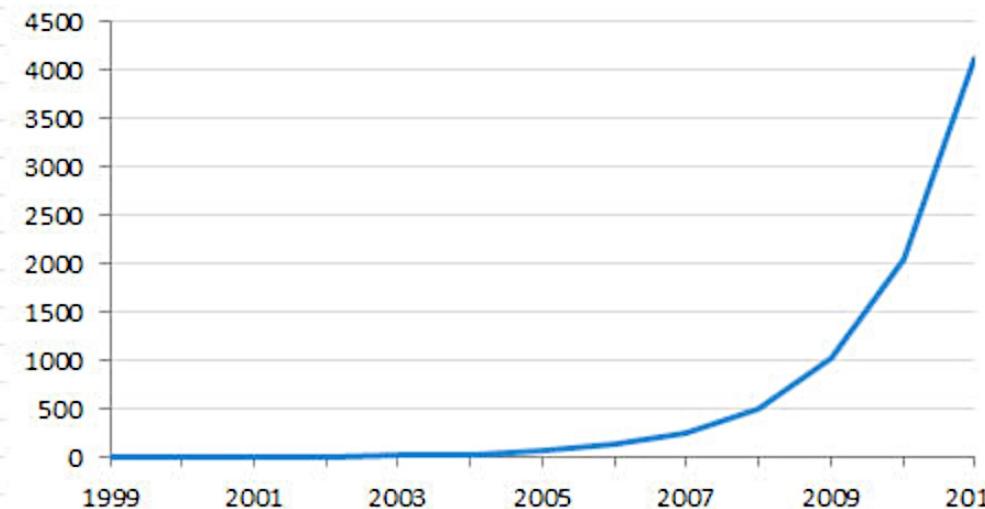


Facilitate Comparison

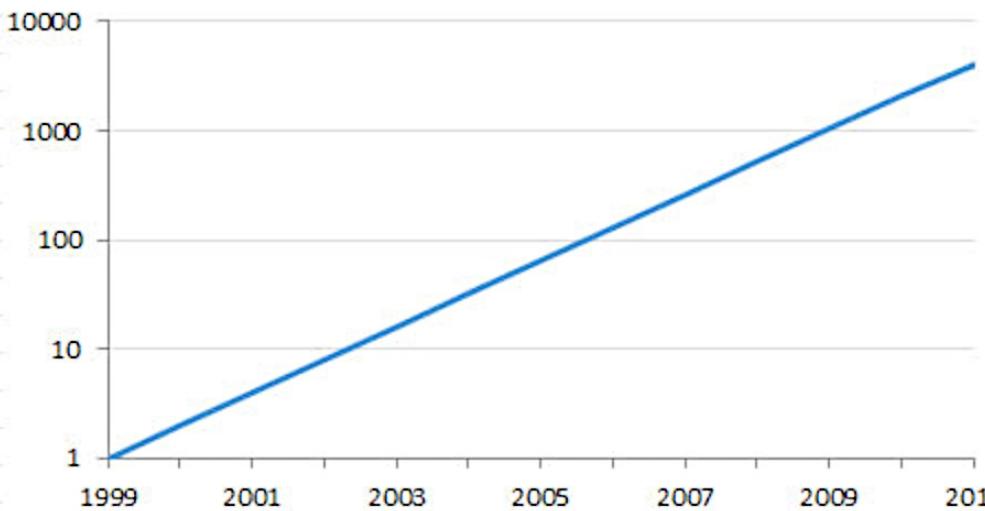
A spaghetti chart of baby names popularity



Linear Scale



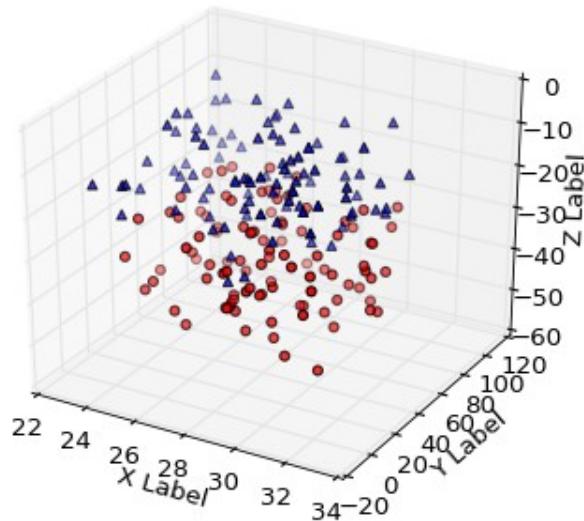
Logarithmic Scale



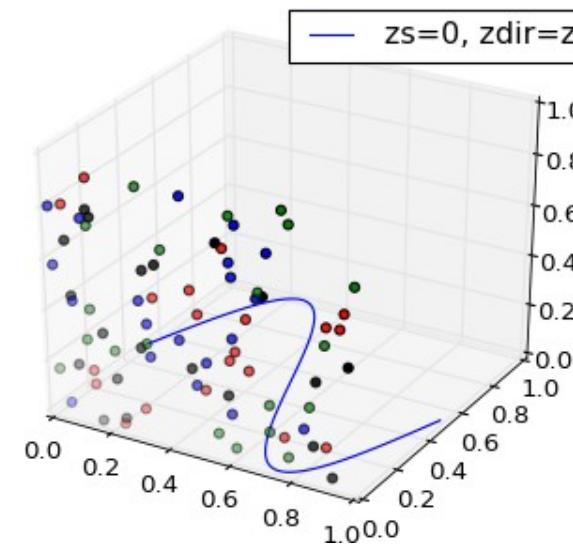
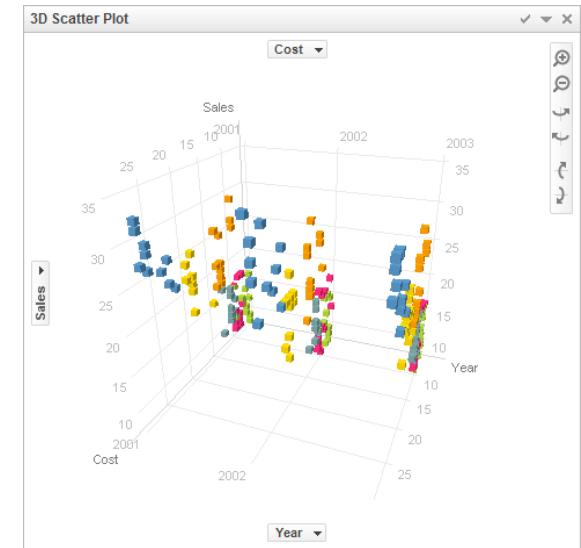
Facilitate Comparison

Consider using a log scale when:

- Data has skewness towards large values; i.e., cases in which one or a few points are much larger than the bulk of the data.
- It is useful to present percent change or multiplicative factors



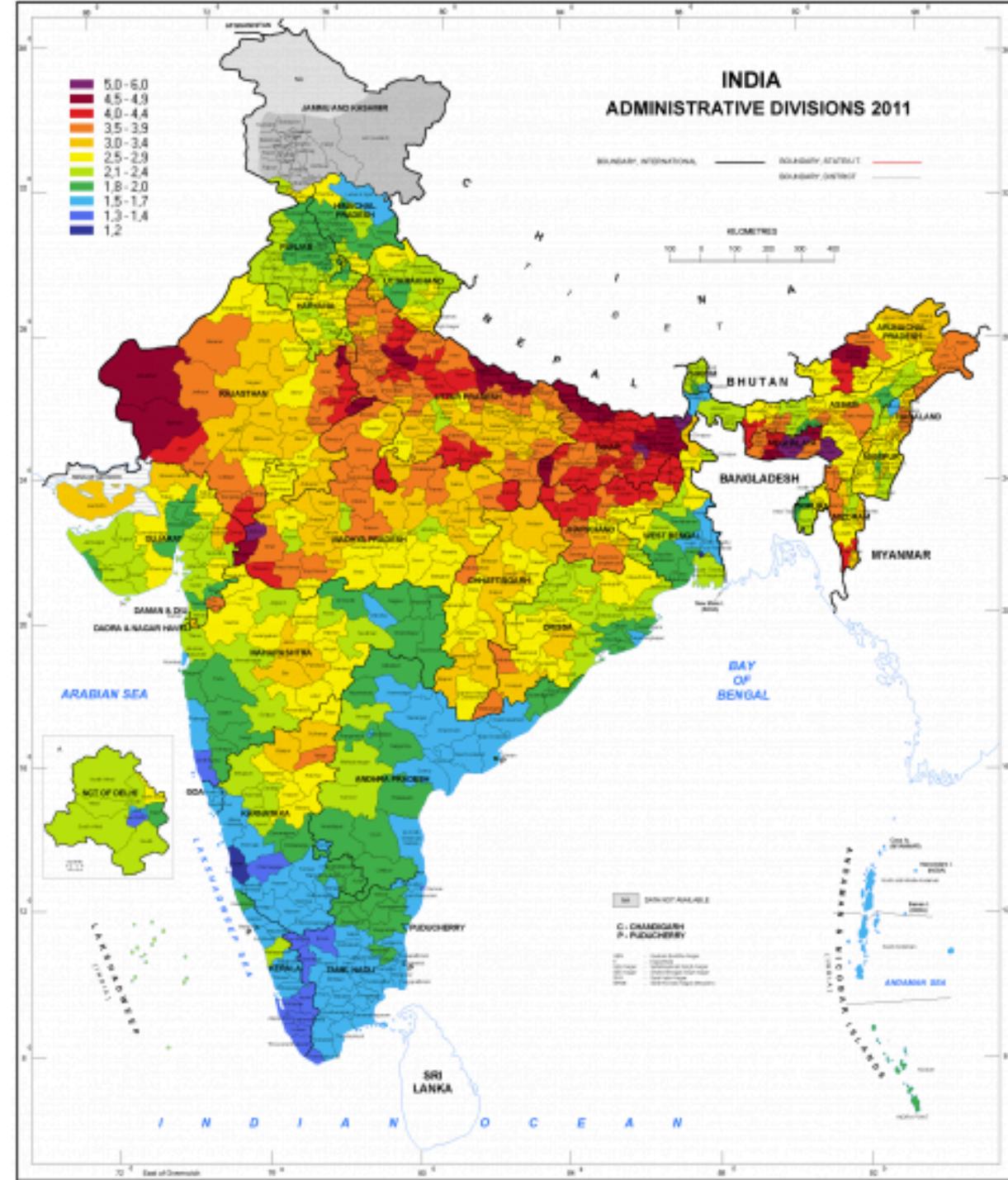
Don't!



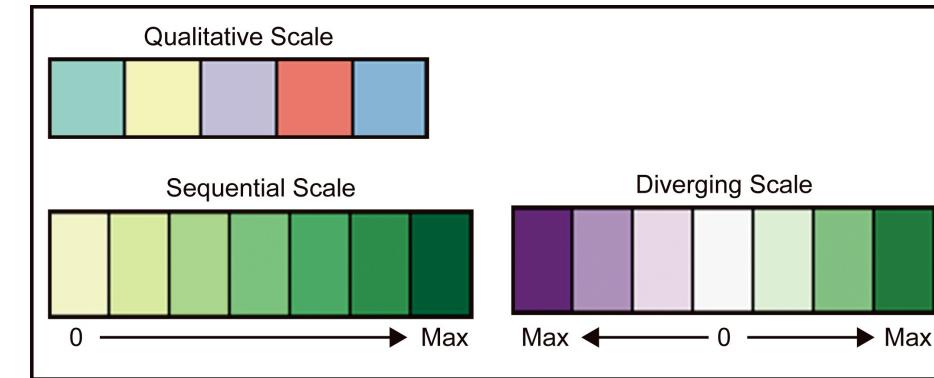
4. Use color strategically

Least Effective

Total fertility rate map: average births per woman by districts, 2011

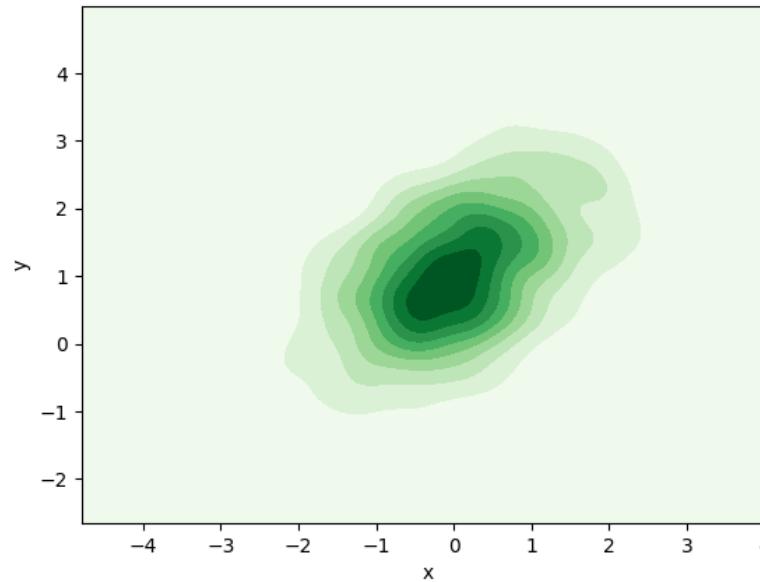


Nominal

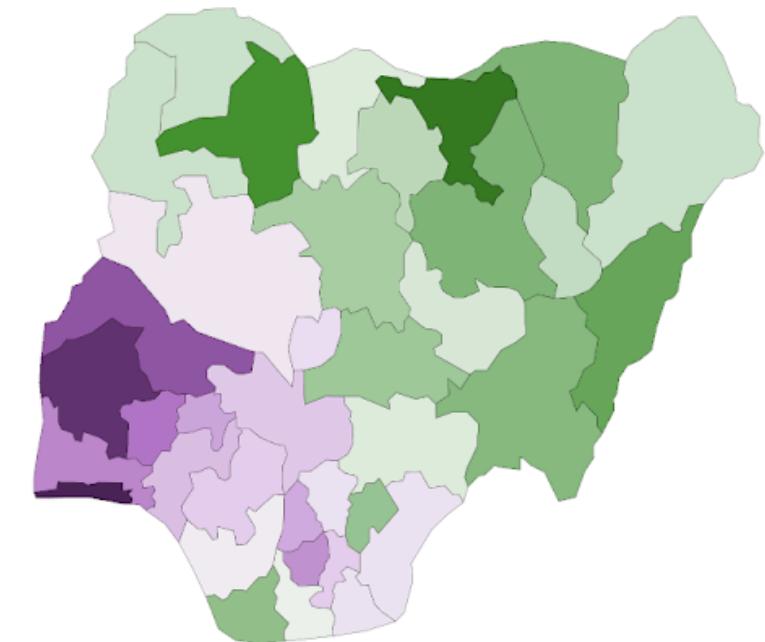


Ordinal

Ex. Densities

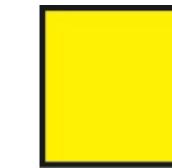


Ex. Correlations



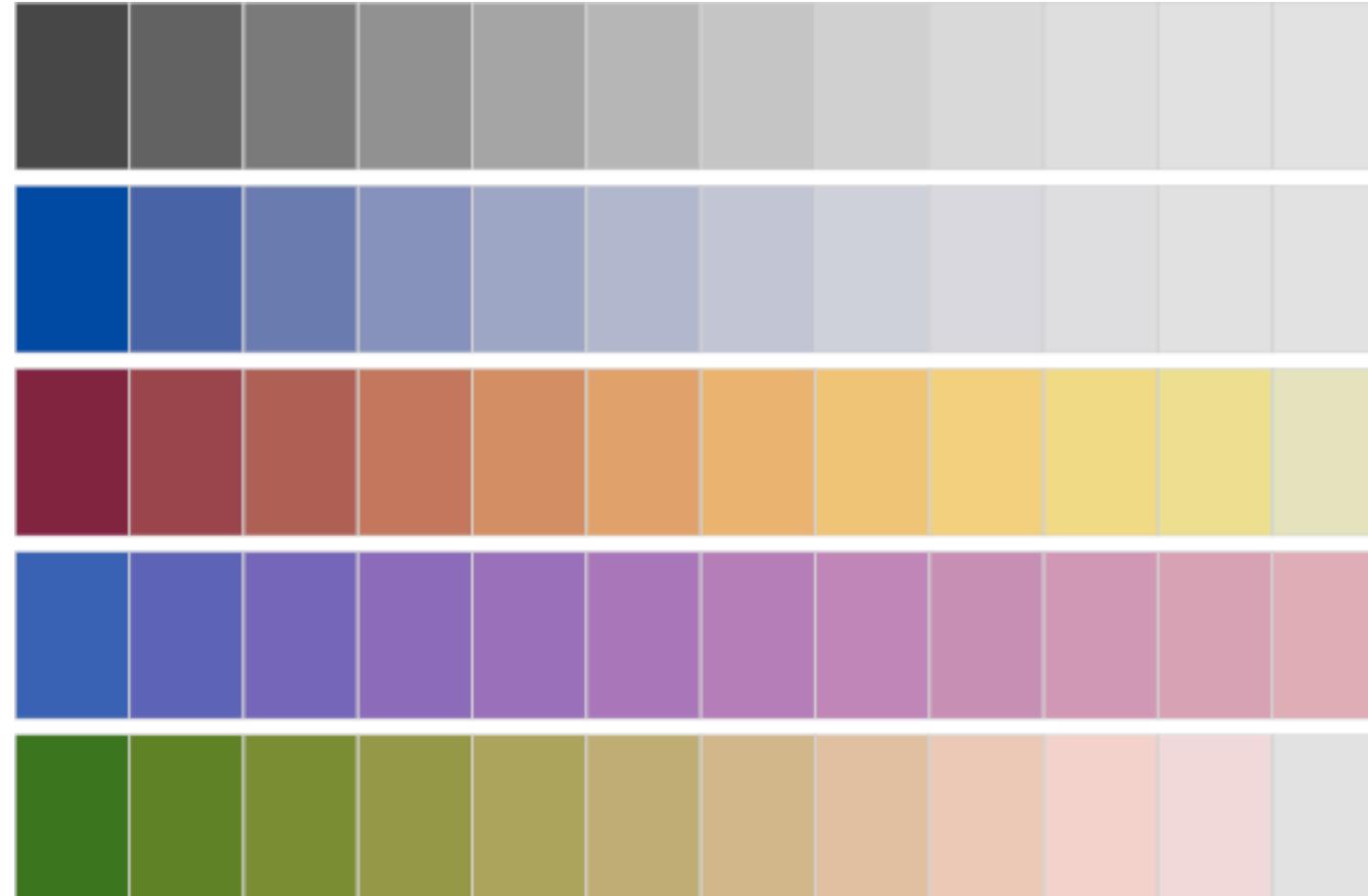
Colours for Categories

Do not use more than 5 colors at once



Colours for Ordinal Data

Vary luminance and saturation



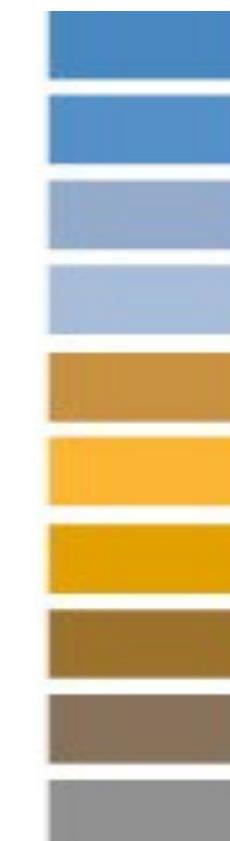
Colour Blindness



Most likely



Protanope



Deutanope

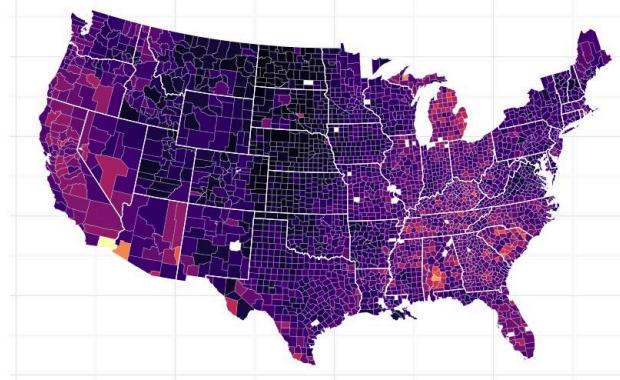


Lightness

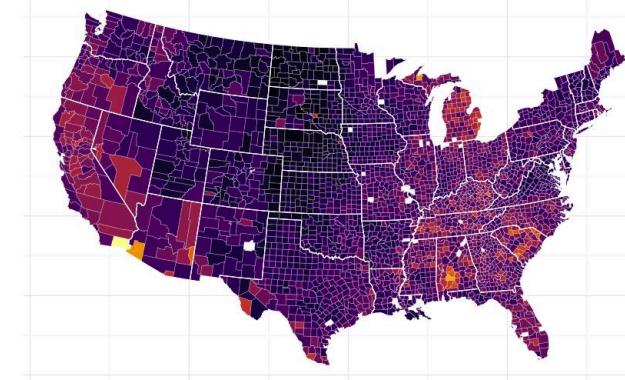
Colourmaps

US unemployment rate by county

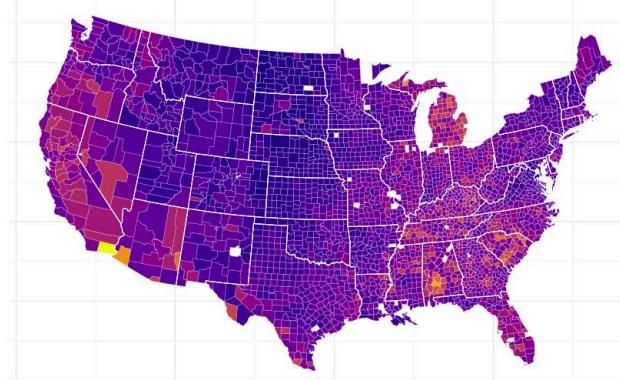
option A aka 'magma'



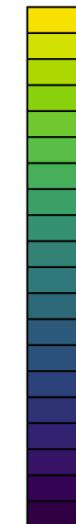
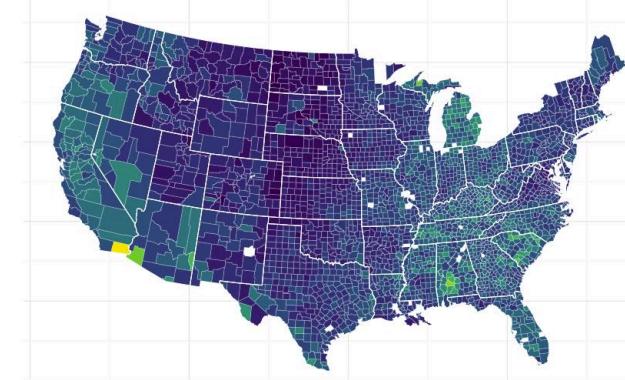
option B aka 'inferno'



option C aka 'plasma'



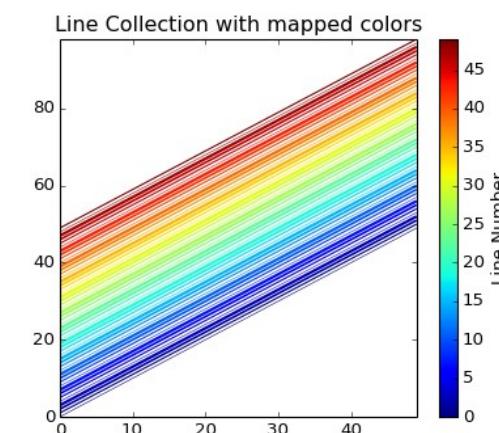
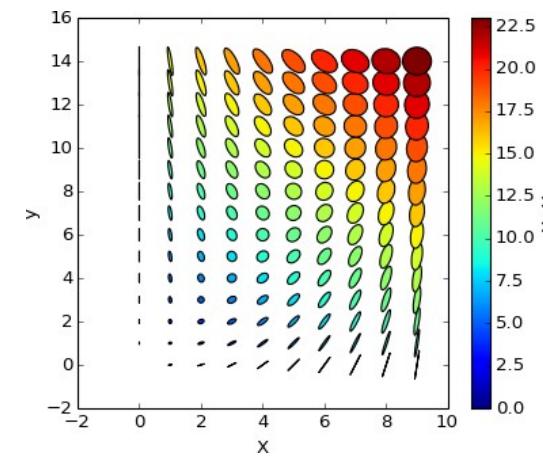
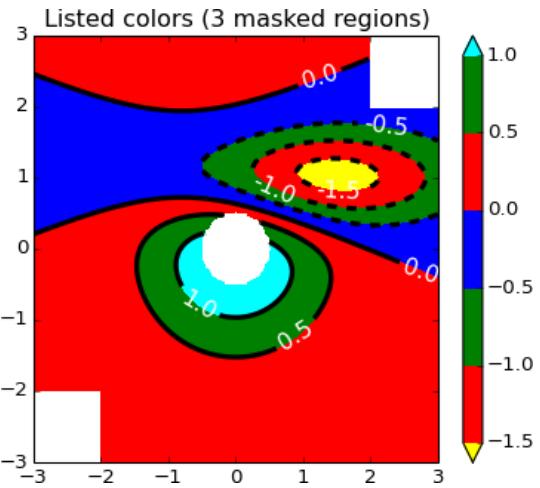
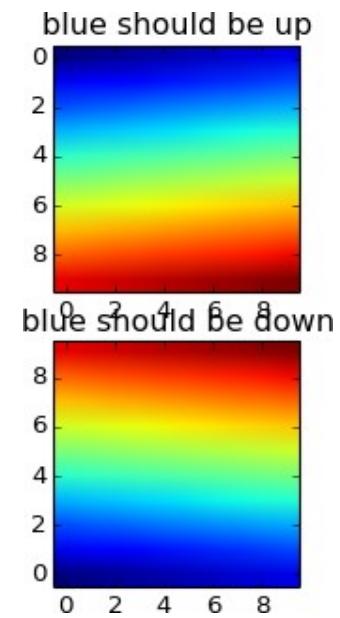
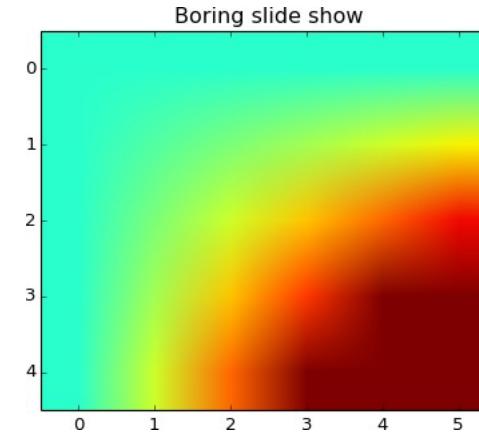
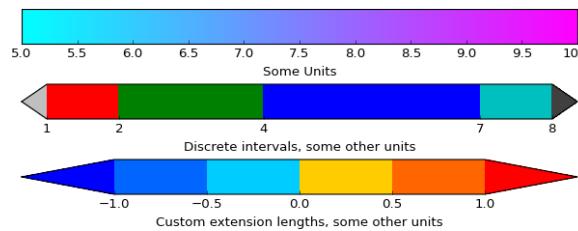
option D aka 'viridis'



Viridis color map for better changes in perception

Avoid Rainbows!

matplotlib gallery



Checklist for Graphical Excellence

(check lecture tab for resources)

- Show the data
- Induce the viewer to think about the substance of the findings rather than the methodology, the graphical design, or other aspects
- Avoid distorting what the data have to say
- Present many numbers in a small space, i.e., efficiently
- Make large datasets coherent
- 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
- Serve a clear purpose: description, exploration, tabulation, or decoration
- Be closely integrated with the statistical and verbal descriptions of the dataset
- Exclude unneeded dimensions
- Omit "chart junk" (term from E.R. Tufte) and unnecessary ink
- Present data in a way to facilitate comparisons
- Make efficient use of space
- Select the best graph type
- Show uncertainty
- Explore several ways to display the data!

For next class..



Finish Lab 04 to practice programming



Submit Homework 04 for peer review on Peer



Check Assignment 2 – due in **Week 5** on Friday at **2330**



See “To do before class” for every lecture (~ 1 hour of self study)



Read paper for **Discussion** session before every Friday



Post questions on the **Discussion** forum on Brightspace (especially on **Visualisation** for this week)