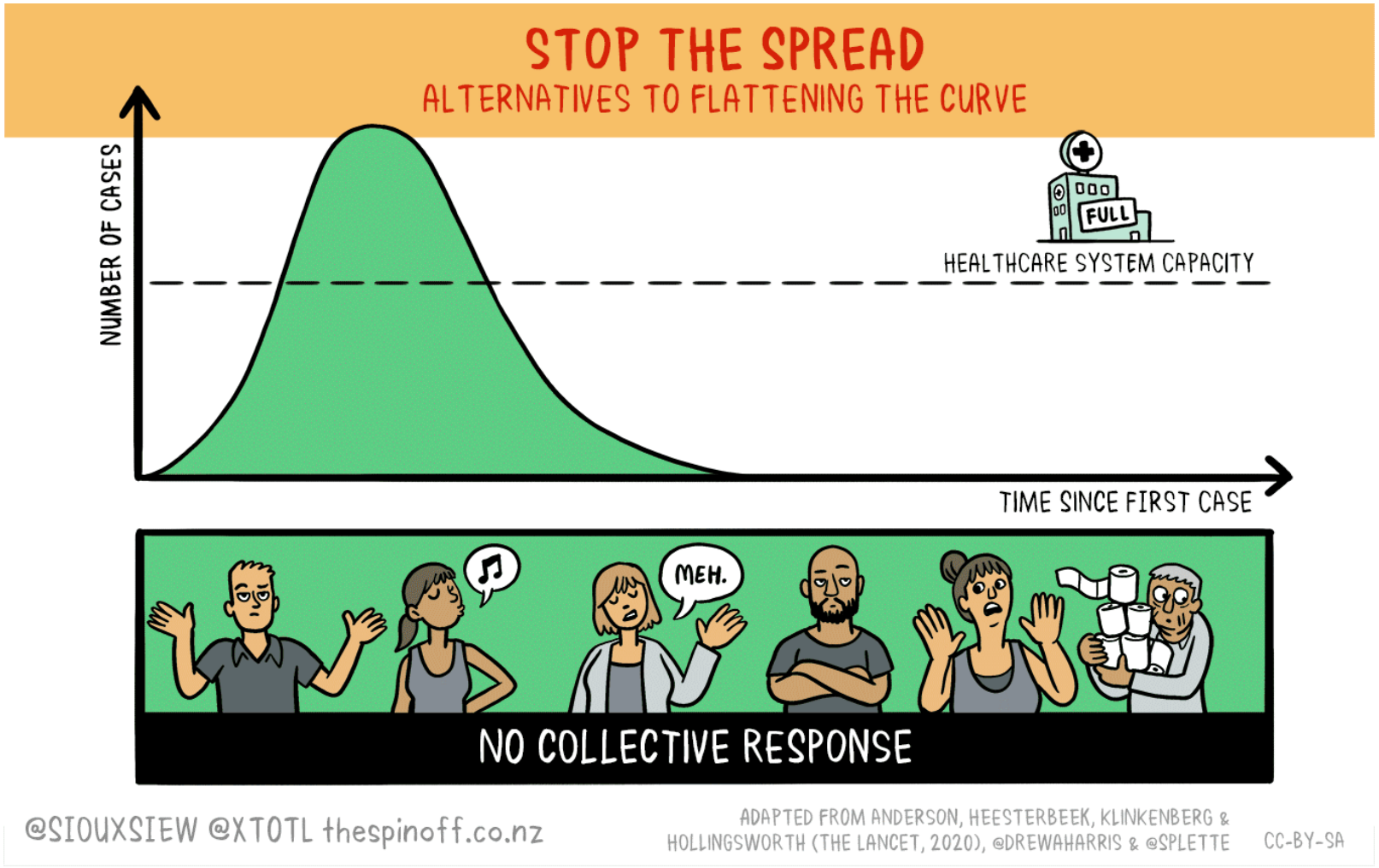
Spatial Data Science

EDA & Visualisation

(EPA122A)

Lecture 5

Trivik Verma



Group formation (Deadline 15 Dec 2023)

Form groups of ***4 students*** each. Each student needs to be part of a group. If you are going to drop the course, don't sign up for the final project as that may delay the progress of other students.

Some suggestions for creating effective groups,

- Be inclusive of people who might not be able to come to class. If they have reached out to your group, find an adequate hybrid arrangement of working together.
- Strive for diverse groups. (Machine Learning and AI suffer heavily from the bias of individuals and communities. Diversity is crucial in meaningful and effective work.)
- The most appealing option is to form a group with friends. However, I urge you to form groups where you can challenge yourselves.
- A group of 3 or 5 students will not be accepted.
- Please email TAs if you haven't found a group.
- We will release instructions about the final project after submission of Assignment 2.

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”

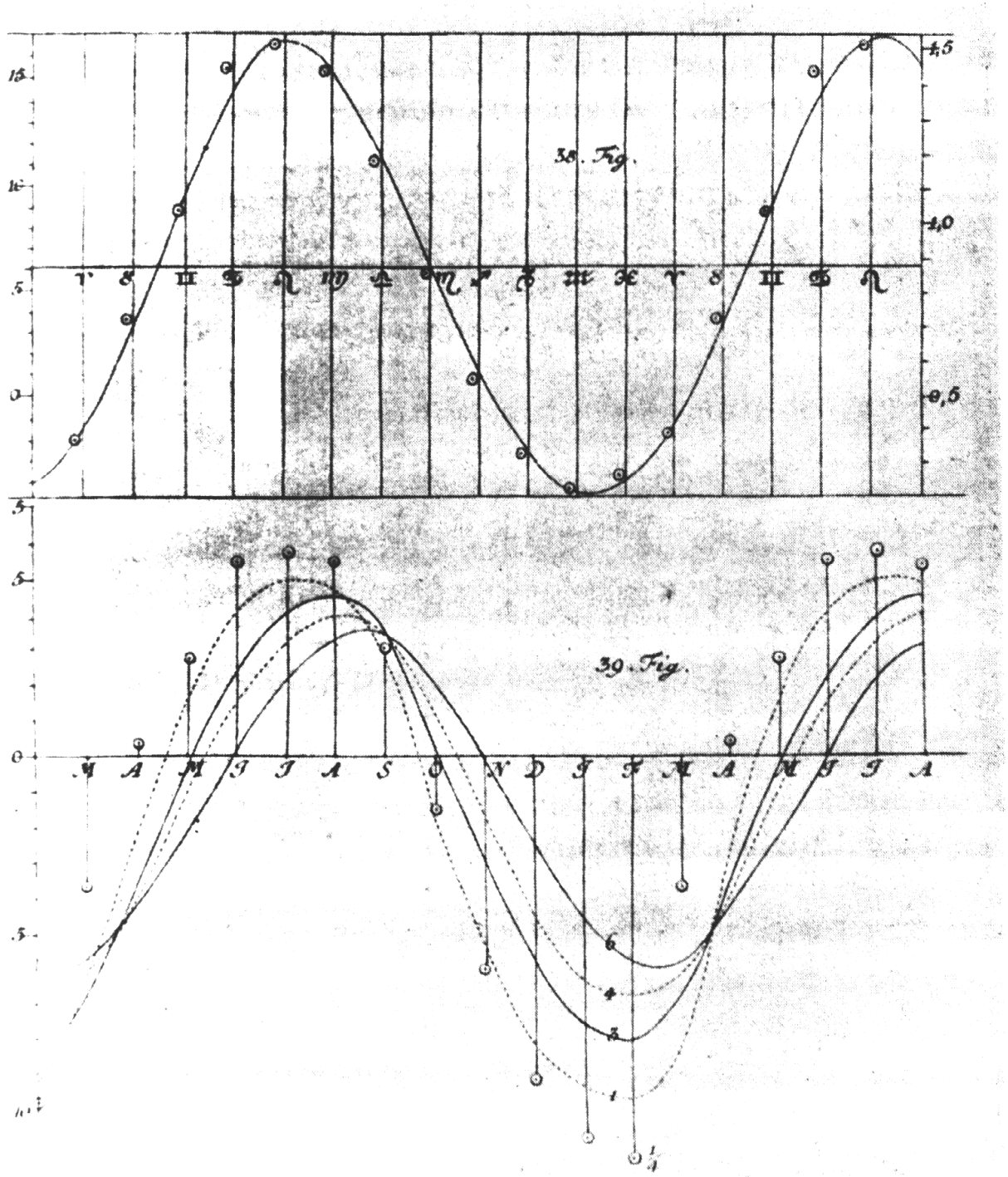
A bit of history

Maps –> Data Maps (17th Century) –> 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.
  - 1786 – Line Plots
  - 1801 – Pie Charts and Circle Graphs
- Other relevant names throughout history:
  Lambert, Minard, Marey

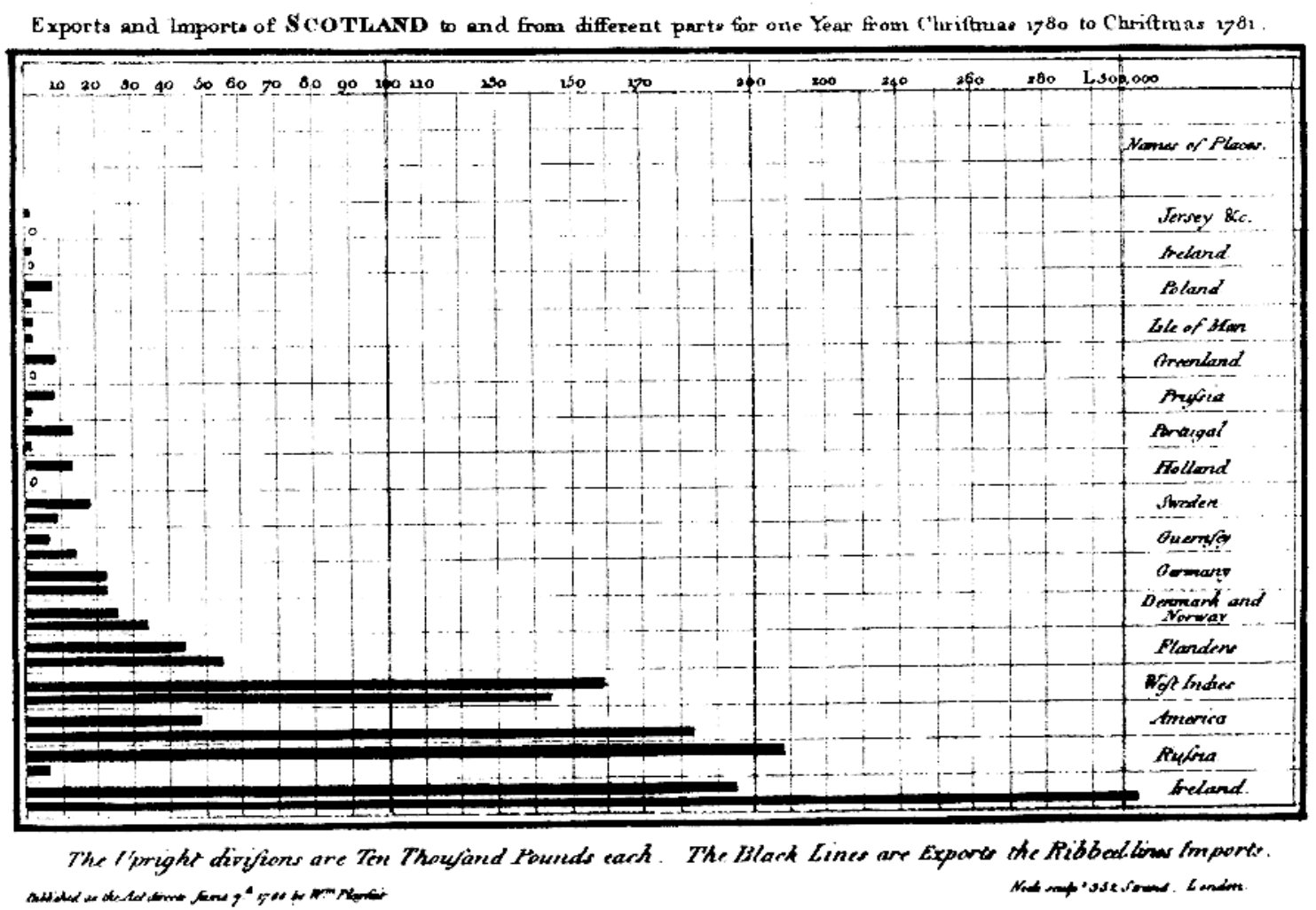
Historical Examples

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



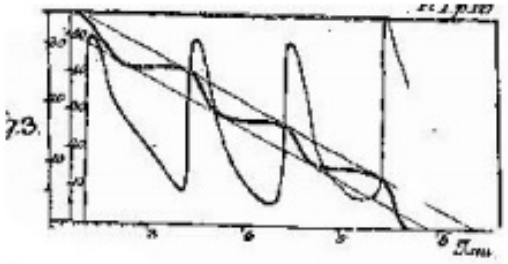
Historical Examples

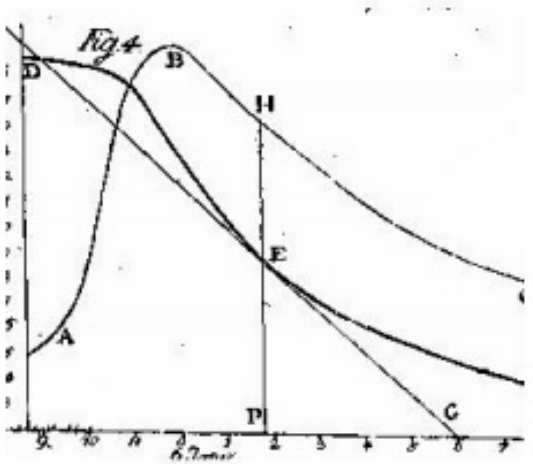
[Source] Playfair 's bar chart in The Commercial and Political Atlas (1786)

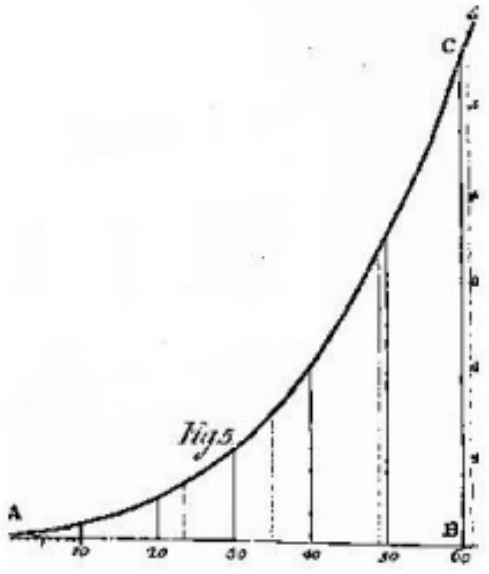


Historical Examples

[Source] Lambert - Evaporation rate against temperature, 1769

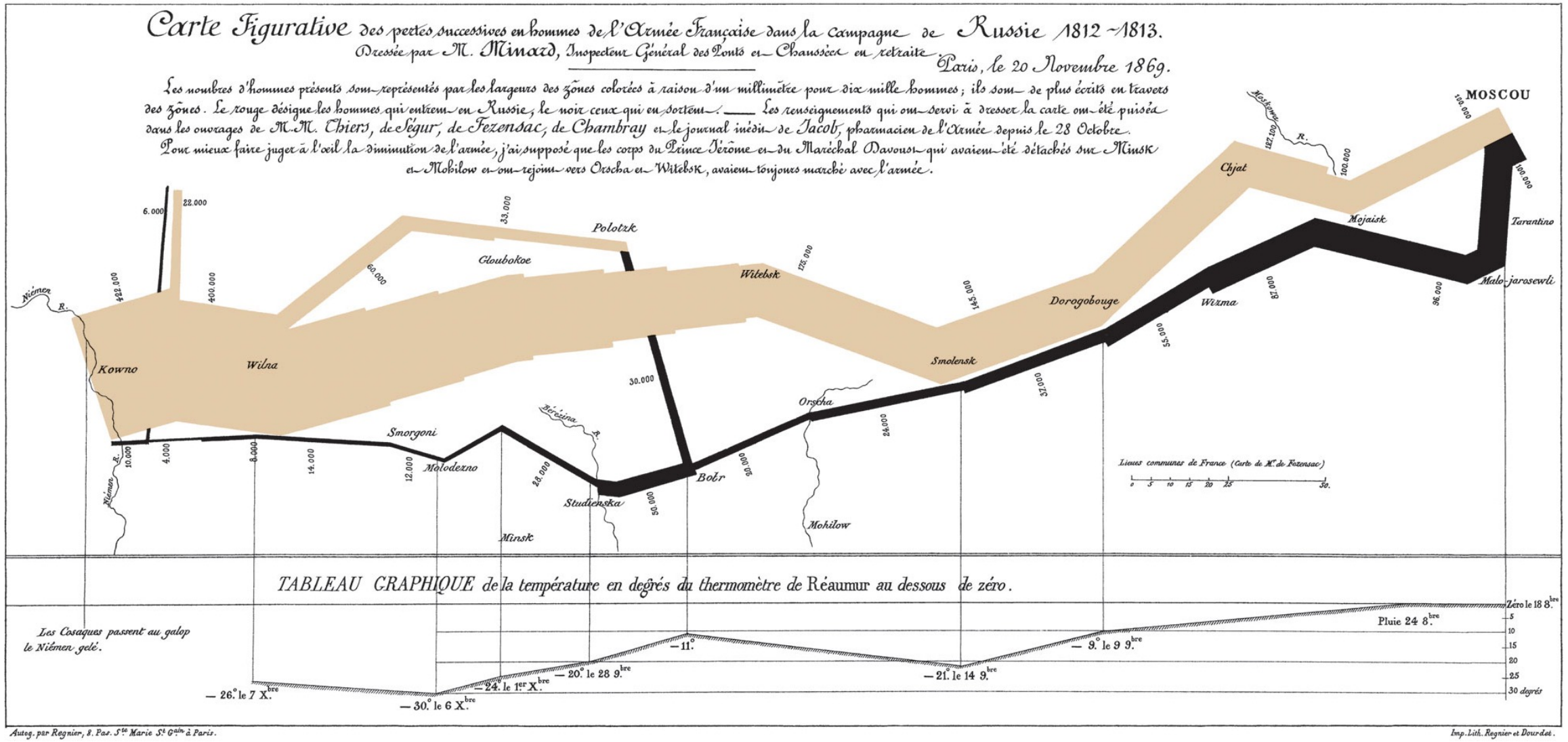






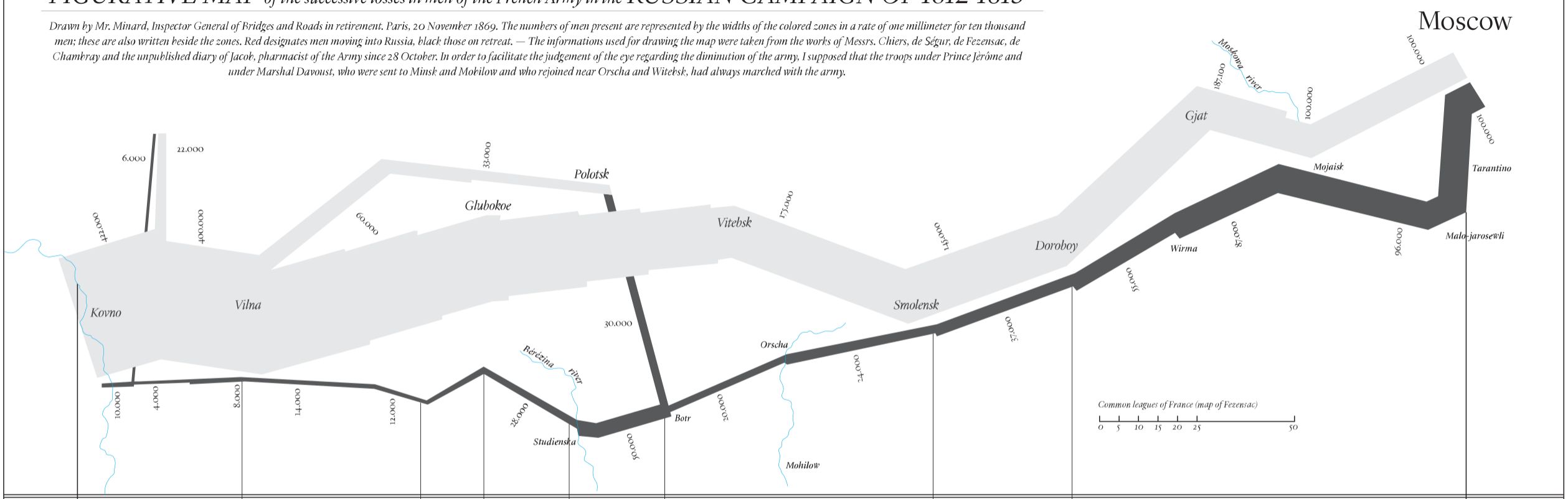
Historical Examples

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

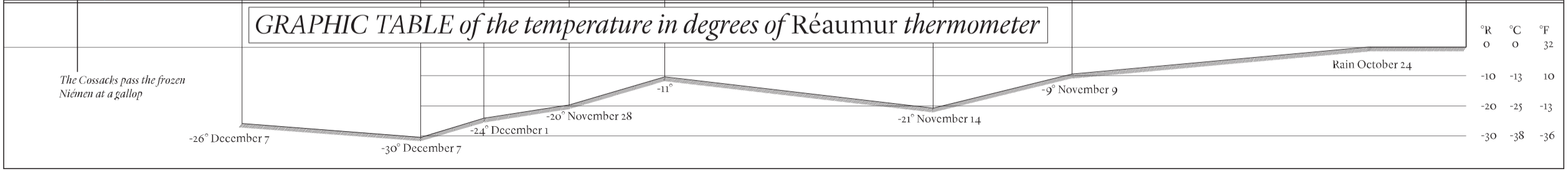


FIGURATIVE MAP of the successive losses in men of the French Army in the RUSSIAN CAMPAIGN OF 1812-1813

Drawn by Mr. Minard, Inspector General of Bridges and Roads in retirement. Paris, 20 November 1869. The numbers of men present are represented by the widths of the colored zones in a rate of one millimeter for ten thousand men; these are also written beside the zones. Red designates men moving into Russia, black those on retreat. — The informations used for drawing the map were taken from the works of Messrs. Chiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, pharmacist of the Army since 28 October. In order to facilitate the judgement of the eye regarding the diminution of the army, I supposed that the troops under Prince Jérôme and under Marshal Davoust, who were sent to Minsk and Mobilow and who rejoined near Orscha and Witebsk, had always marched with the army.

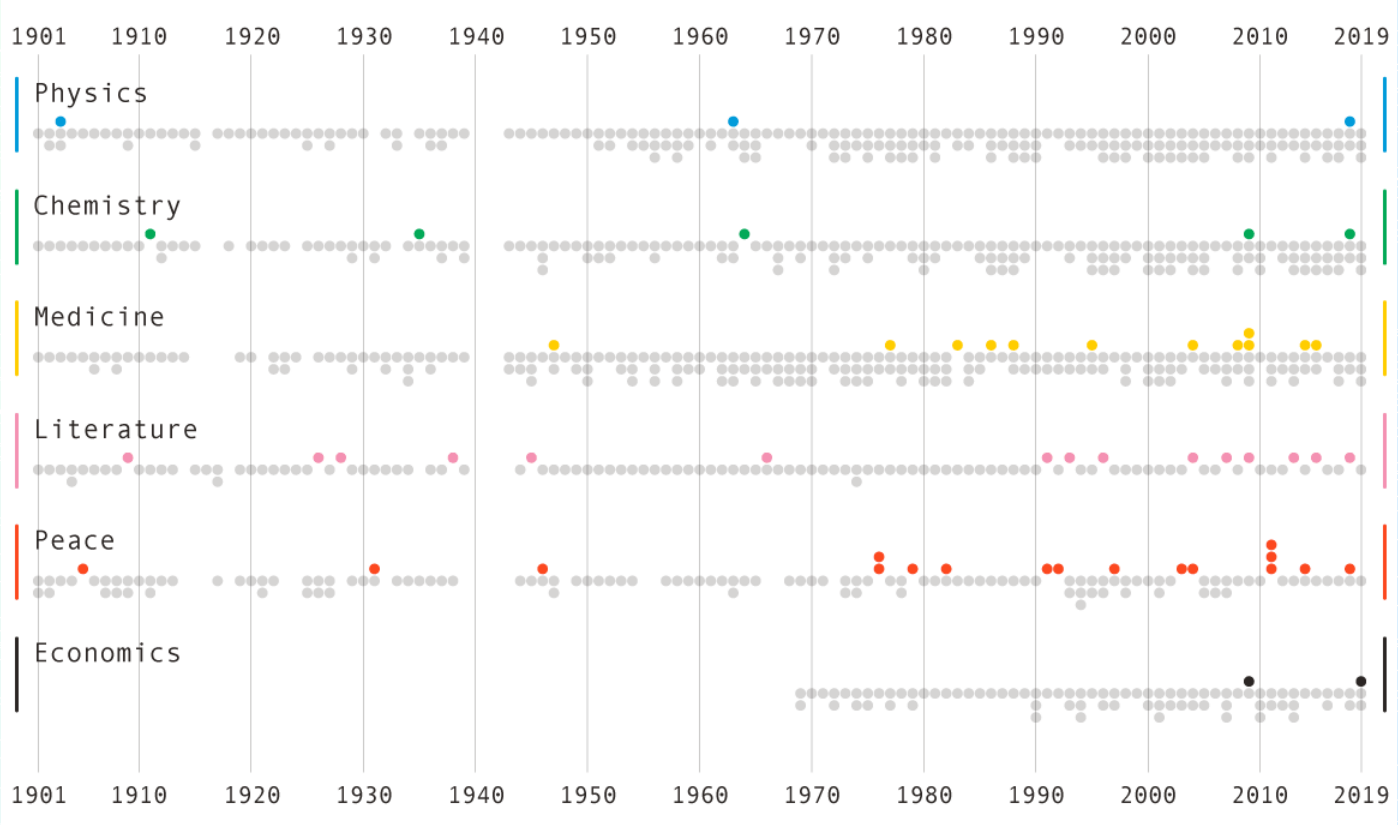


GRAPHIC TABLE of the temperature in degrees of Réaumur *thermometer*



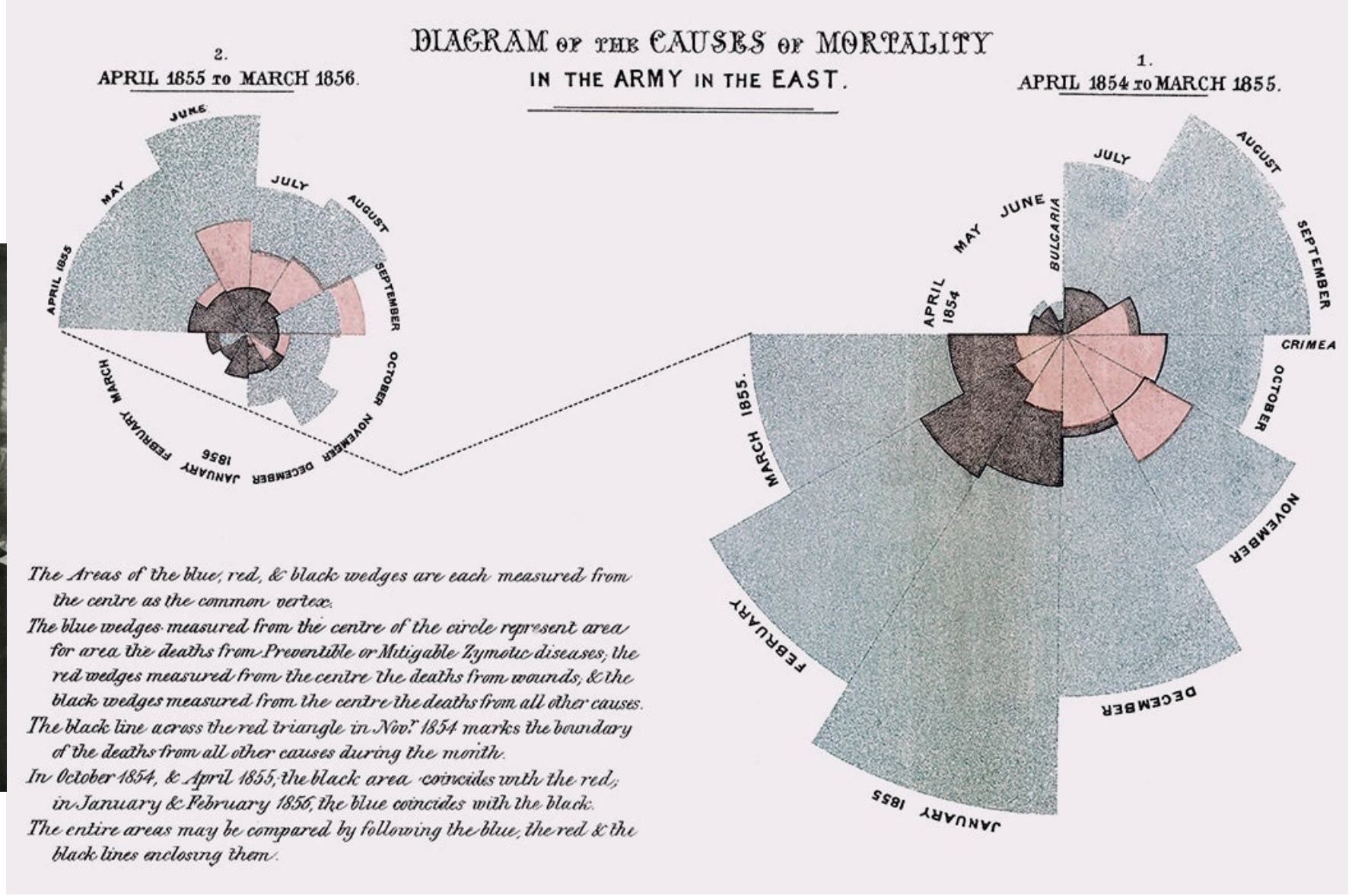
WOMEN NOBEL PRIZE WINNERS

From 1901 to 2019



UN WOMEN

SOURCE | The Nobel Foundation



How Florence Nightingale Changed Data Visualization Forever

https://www.scientificamerican.com/article/how-florence-nightingale-changed-data-visualization-forever/

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)

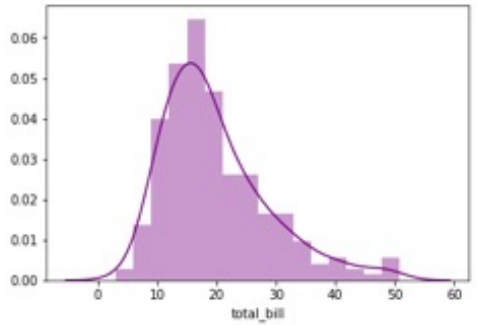
- Explore the data
- Assess a situation
- Identify hidden patterns and trends
- Formulate/test hypothesis
- Decide what to do next in analysis/modelling

Communicate (Explanatory)

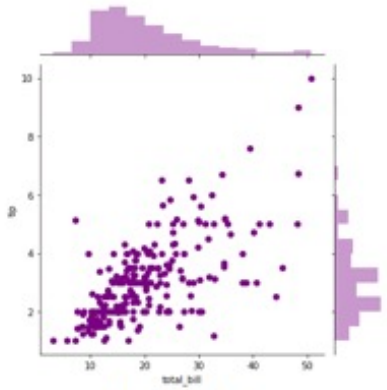
- Present information and ideas succinctly
- Explain and inform
- Provide evidence and support
- Influence and persuade

Visualisation Goals

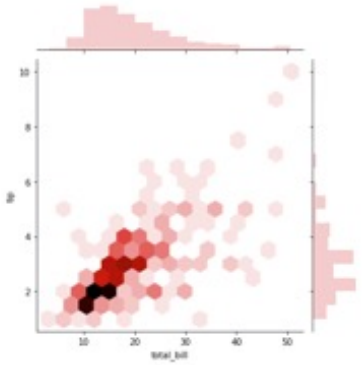
Analyse (Exploratory)



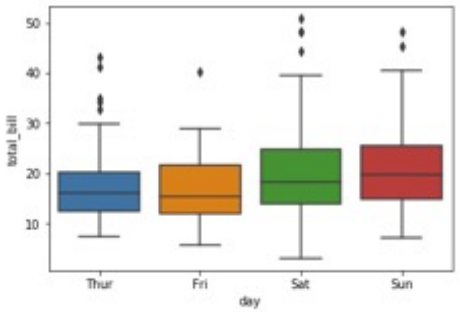
distplot



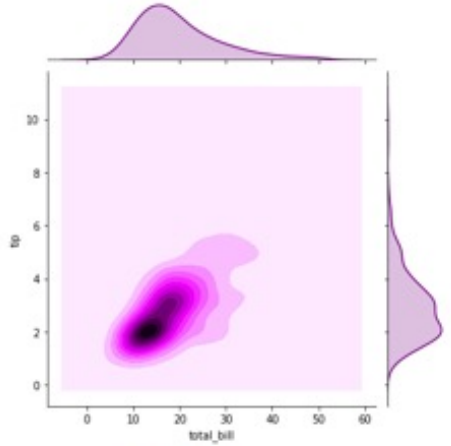
Jointplot



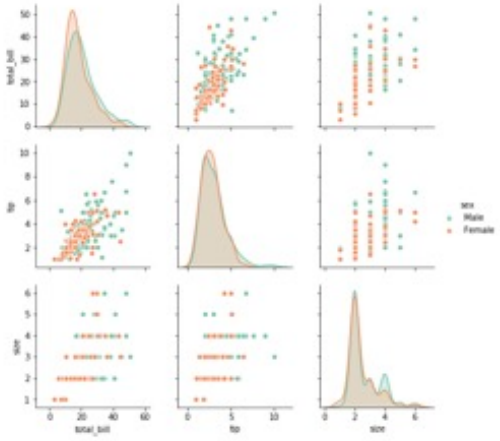
Hexplots



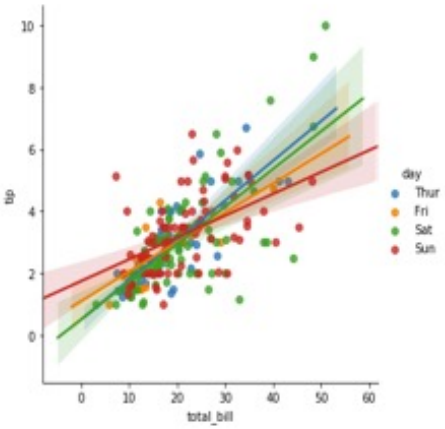
Boxplots



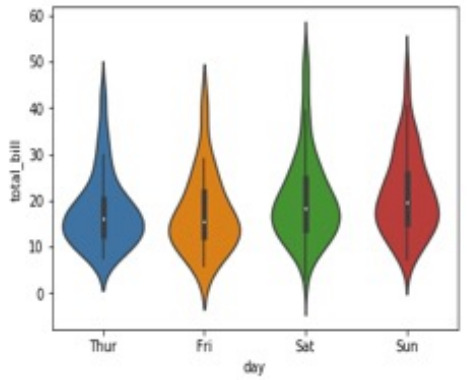
KDE Plot



Pair Plots



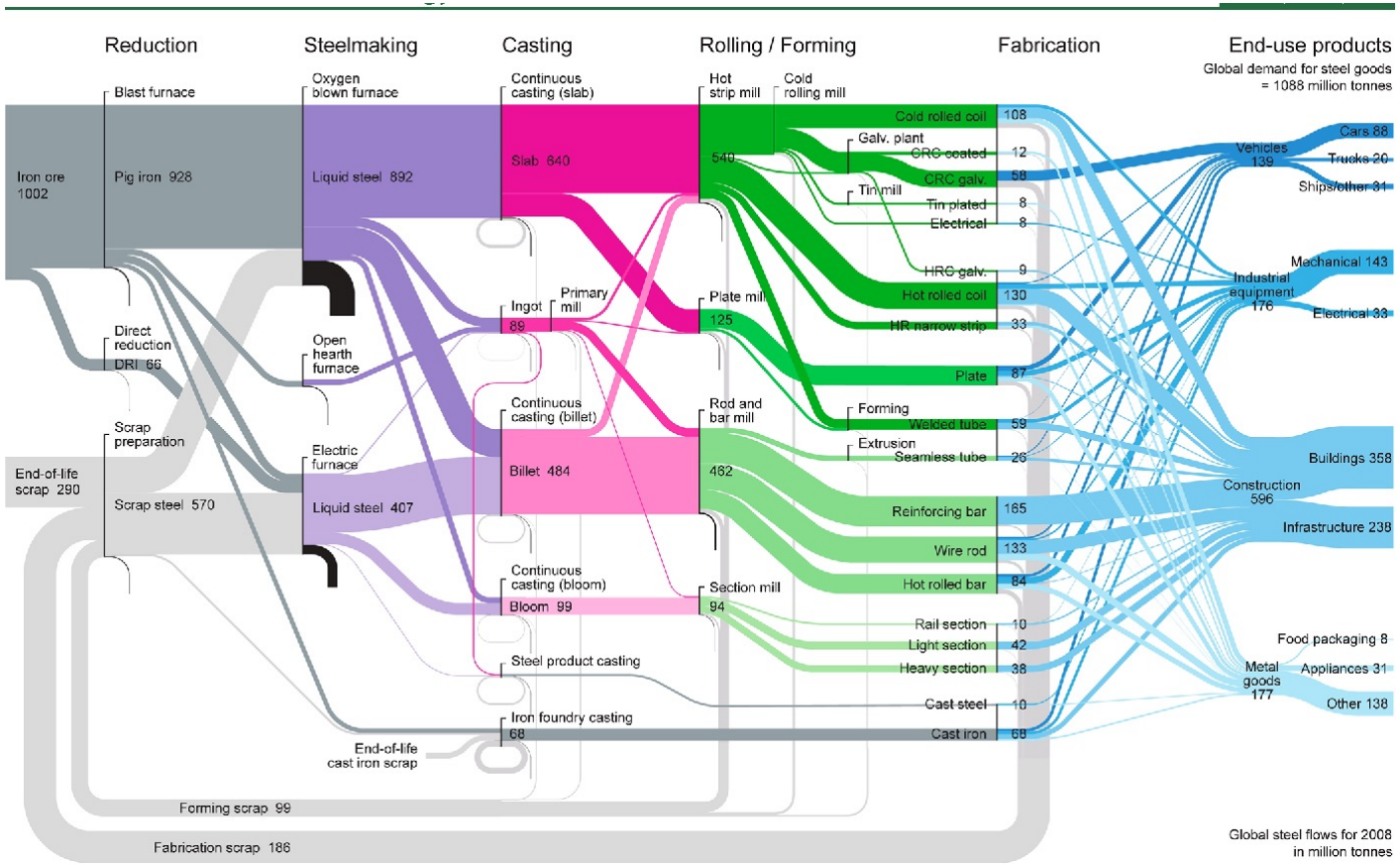
LM Plots



Violin Plots

Visualisation Goals

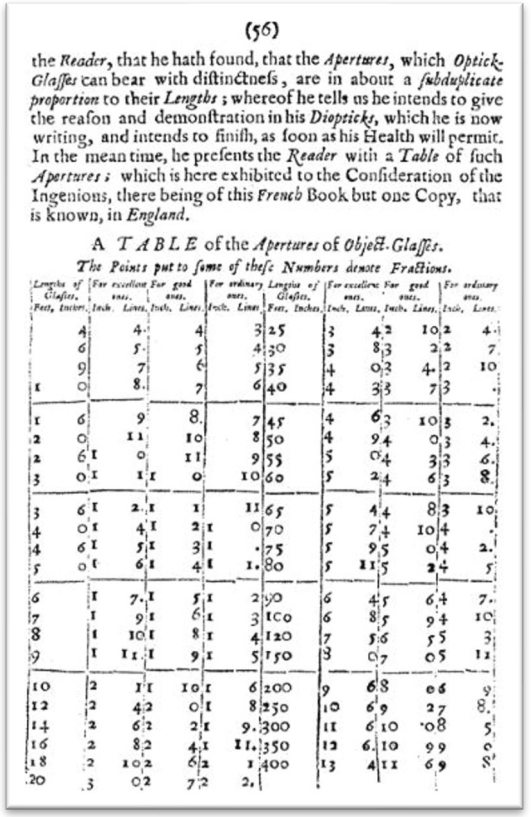
Communicate (Explanatory)



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

Cullen, J. M., Allwood, J. M., & Bambach, M. D. (2012). Mapping the global flow of steel: from steelmaking to end-use goods. *Environmental science & technology*, *46*(24), 13048-13055.

Tables vs Graphs



(56)

the *Reader*, that he hath found, that the *Apertures*, which *Optick-Glaſſes* can bear with diſtinctneſs, are in about a *ſubduplicate proportion* to their *Lengths*; whereof he tells us he intends to give the reaſon and demonſtration in his *Diopticks*, 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*.

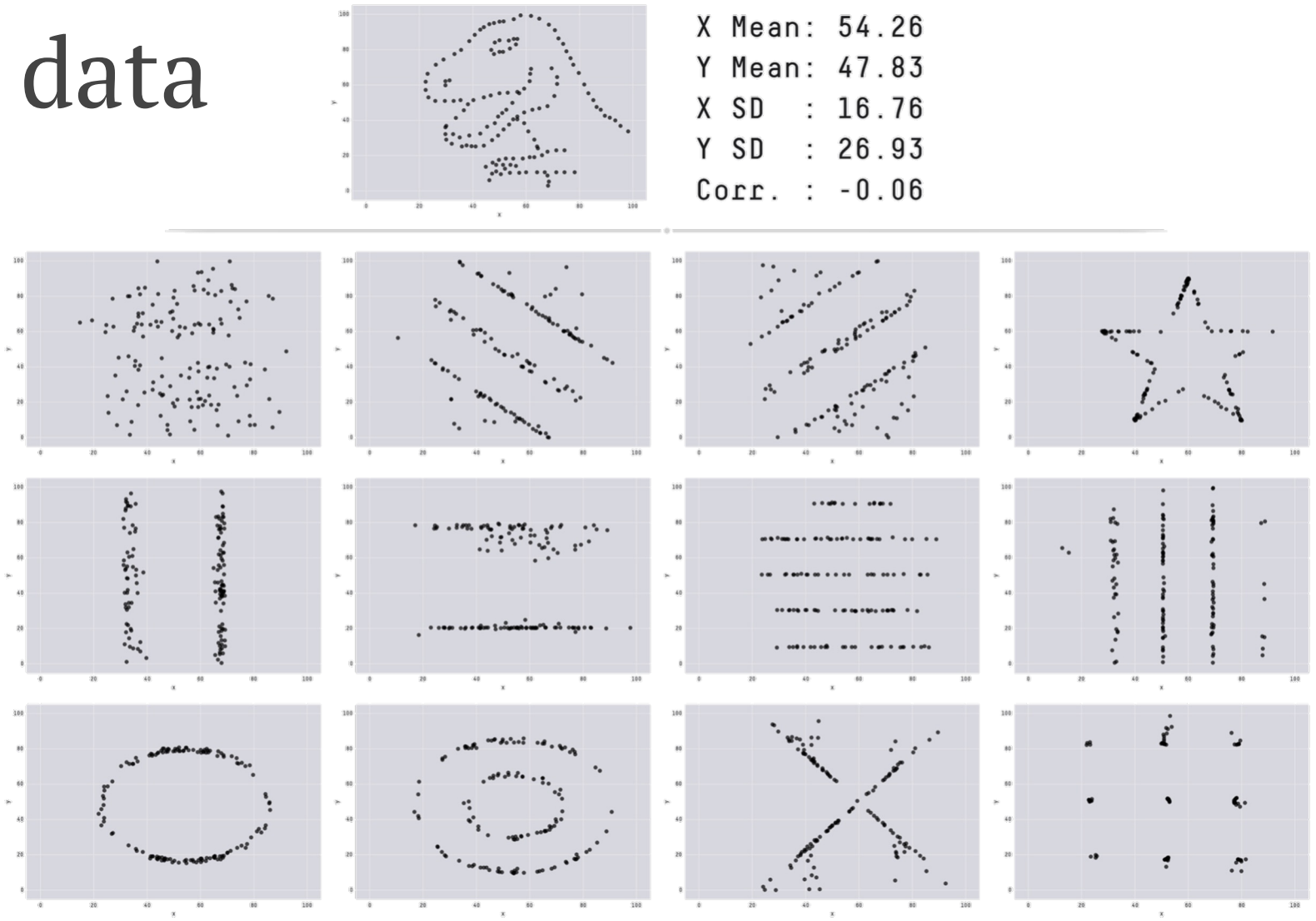
The Points put to ſome of theſe Numbers denote Fractions.

Lengths of Glaſſes. Feet, Inches.	For excellent eyes. Inch. Lines.	For good eyes. Inch. Lines.	For ordinary eyes. Inch. Lines.	Lengths of Glaſſes. Feet. Inches.	For excellent eyes. Inch. Lines.	For good eyes. Inch. Lines.	For ordinary eyes. Inch. Lines.
4	4	4	3 25	3	4 2	10 2	4
6	5	5	4 30	3	8 3	2 2	7
9	7	6	5 35	4	0 3	4 2	10
1 0	8	7	6 40	4	3 3	7 3	
1 6	9	8	7 45	4	6 3	10 3	2
2 0	11	10	8 50	4	9 4	0 3	4
2 6 1	0	11	9 55	5	0 4	3 3	6
3 0 1	1 1	0	10 60	5	2 4	6 3	8
3 6 1	2 1	1	11 65	5	4 4	8 3	10
4 0 1	4 1	2 1	0 70	5	7 4	10 4	
4 6 1	5 1	3 1	75	5	9 5	0 4	2
5 0 1	6 1	4 1	1 80	5	11 5	2 4	5
6 1	7 1	5 1	2 90	6	4 5	6 4	7
7 1	9 1	6 1	3 100	6	8 5	9 4	10
8 1	10 1	8 1	4 120	7	5 6	5 5	3
9 1	11 1	9 1	5 150	8	0 7	0 5	11
10 2	1 1	10 1	6 200	9	6 8	0 6	9
12 2	4 2	0 1	8 250	10	6 9	2 7	8
14 2	6 2	2 1	9 300	11	6 10	0 8	5
16 2	8 2	4 1	11 350	12	6 10	9 9	0
18 2	10 2	6 2	1 400	13	4 11	6 9	8
20 3	0 2	7 2	2				

- 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

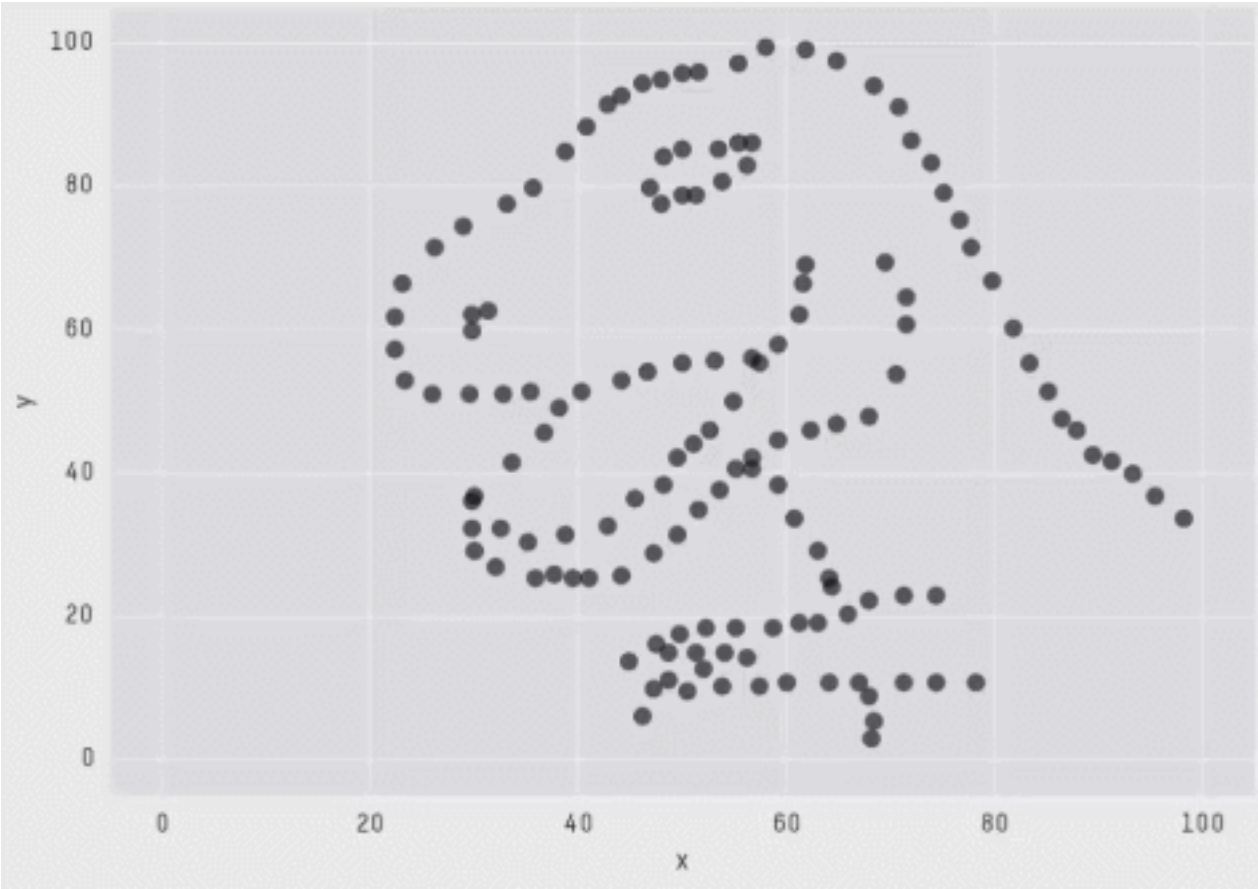


Matejka, J., & Fitzmaurice, G. (2017, May). Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 1290-1294).

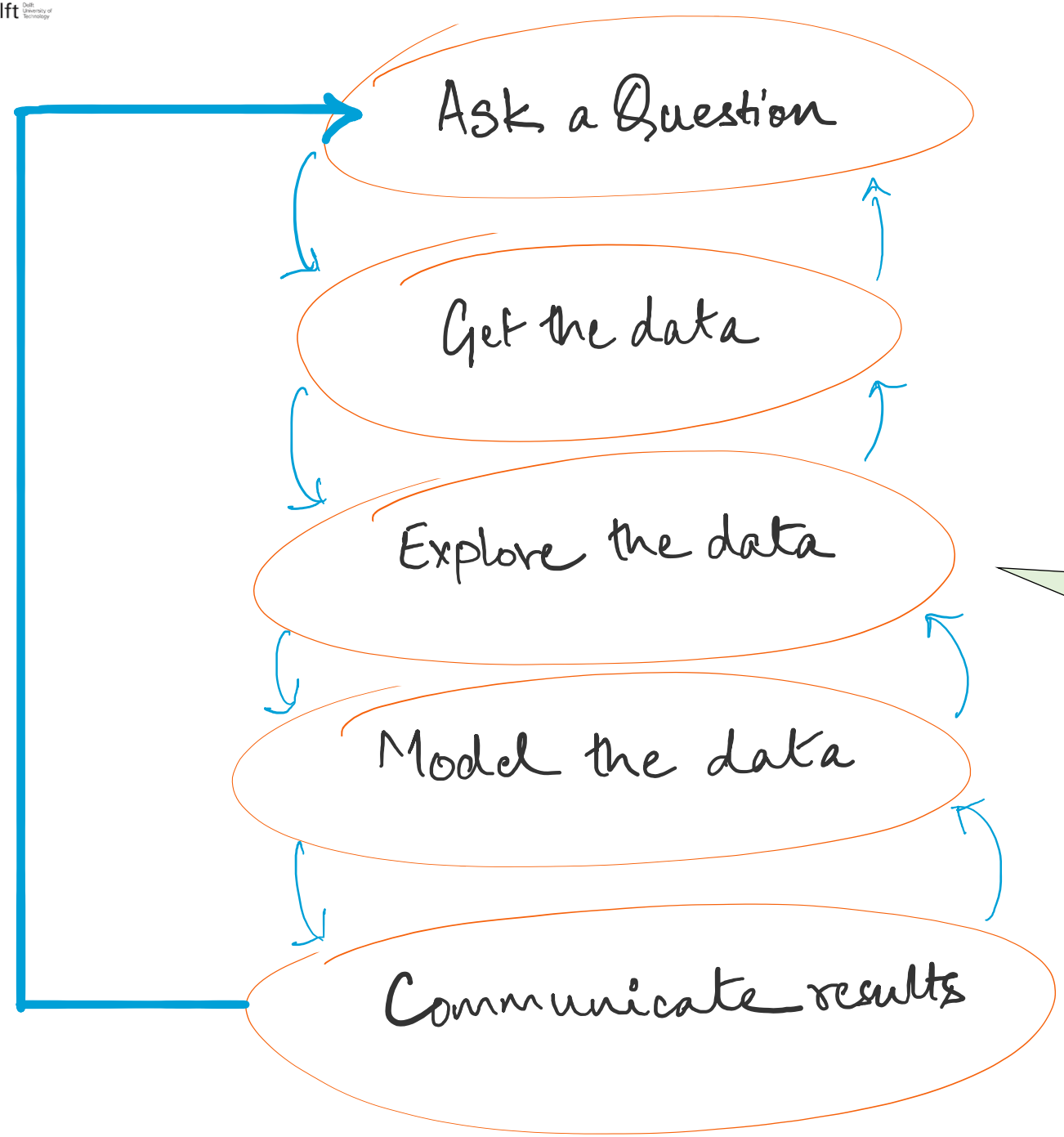
Graphics *reveal* data

Anscombe's quartet

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



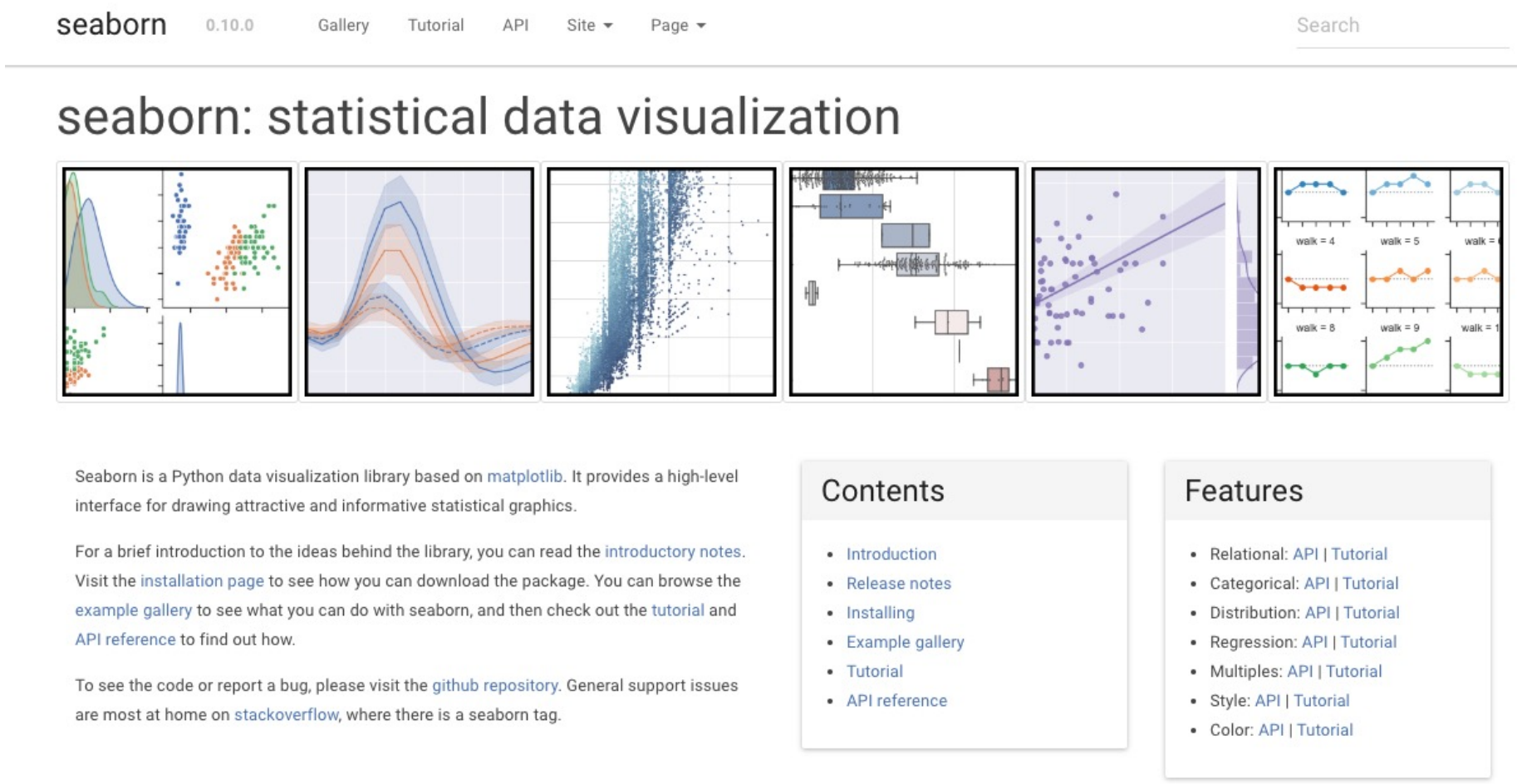
X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526



Plot the data.
Are there **anomalies**?
Are there **patterns**?

Exploratory Data Analysis (EDA)

To convey information through graphical representations of data



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

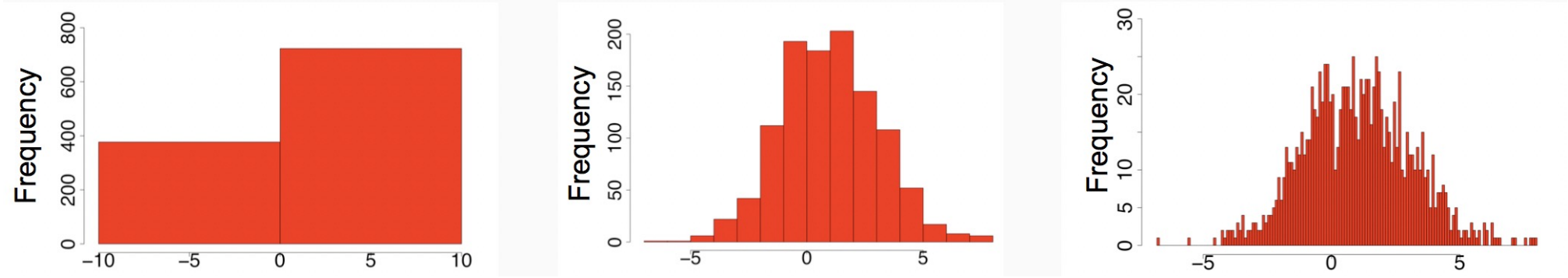
Types of Visualisations

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

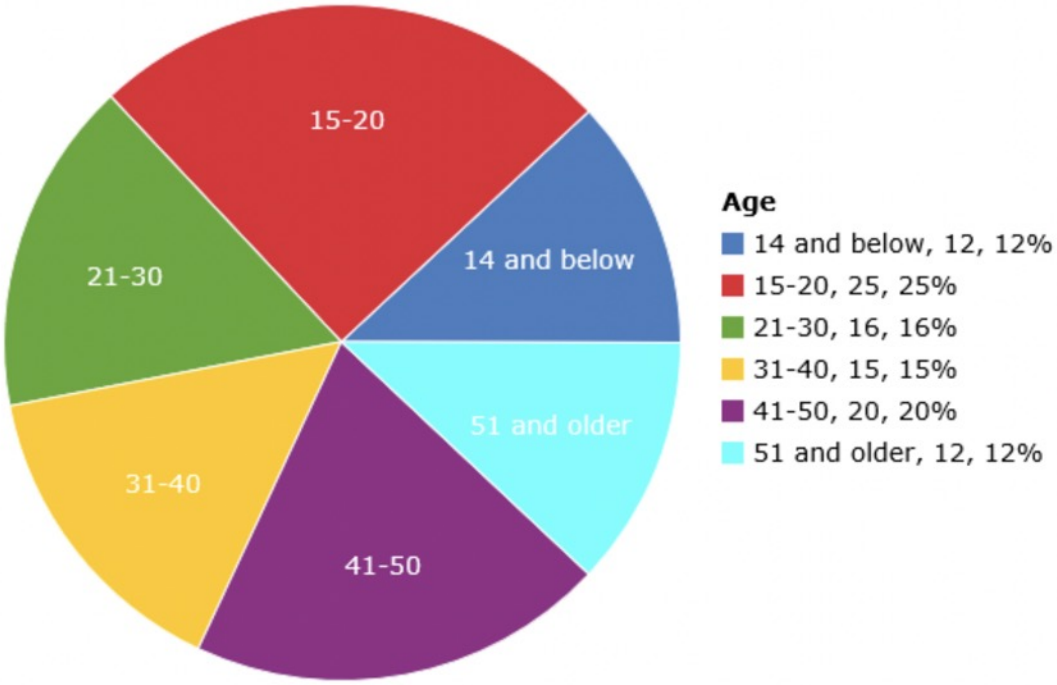
A **histogram** is a way to visualize how 1-dimensional data is distributed across certain values.



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

Pie Charts for Categorical Variables

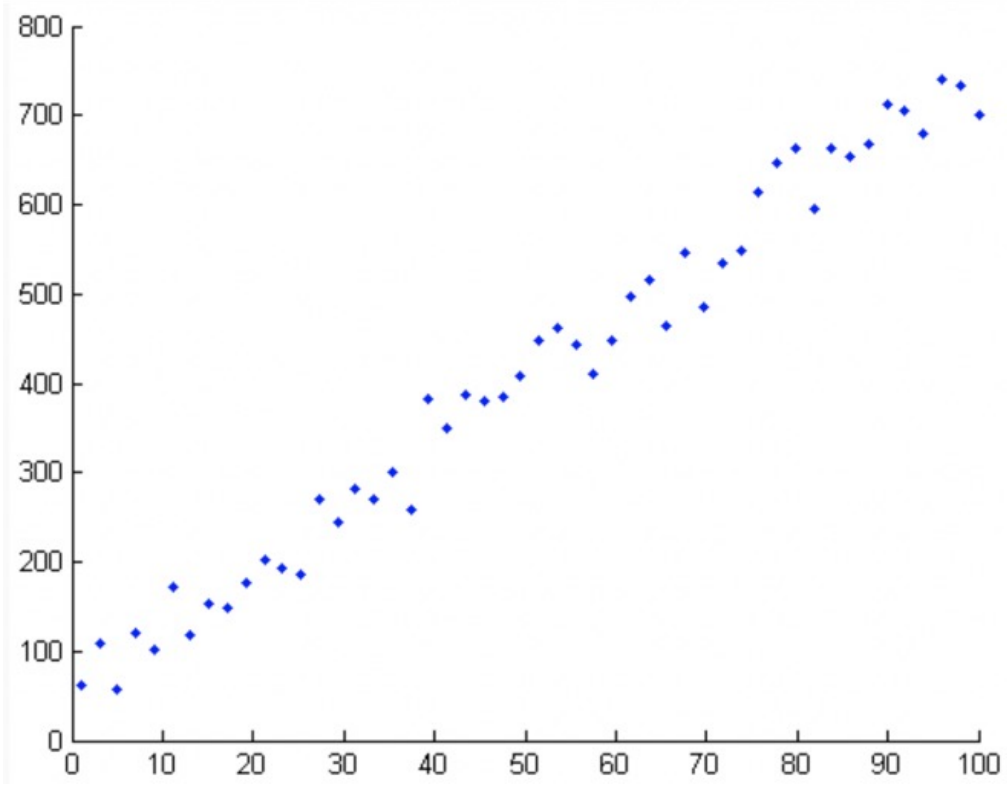
A **pie chart** is a way to visualize the static composition (aka, distribution) of a variable (or single group).



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

Scatter Plots to Visualise Relationships

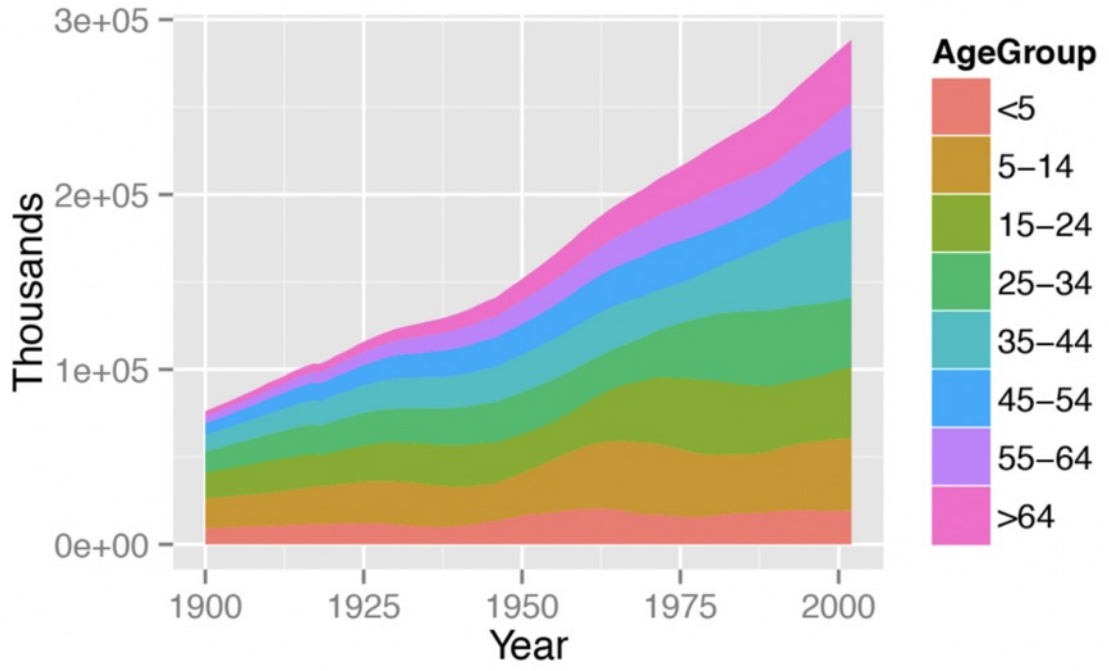
A **scatter plot** is a way to visualize the relationship between two different attributes of multi-dimensional data.



Material provided by the Data Science course at Harvard University

Stacked area graph to show trend over time

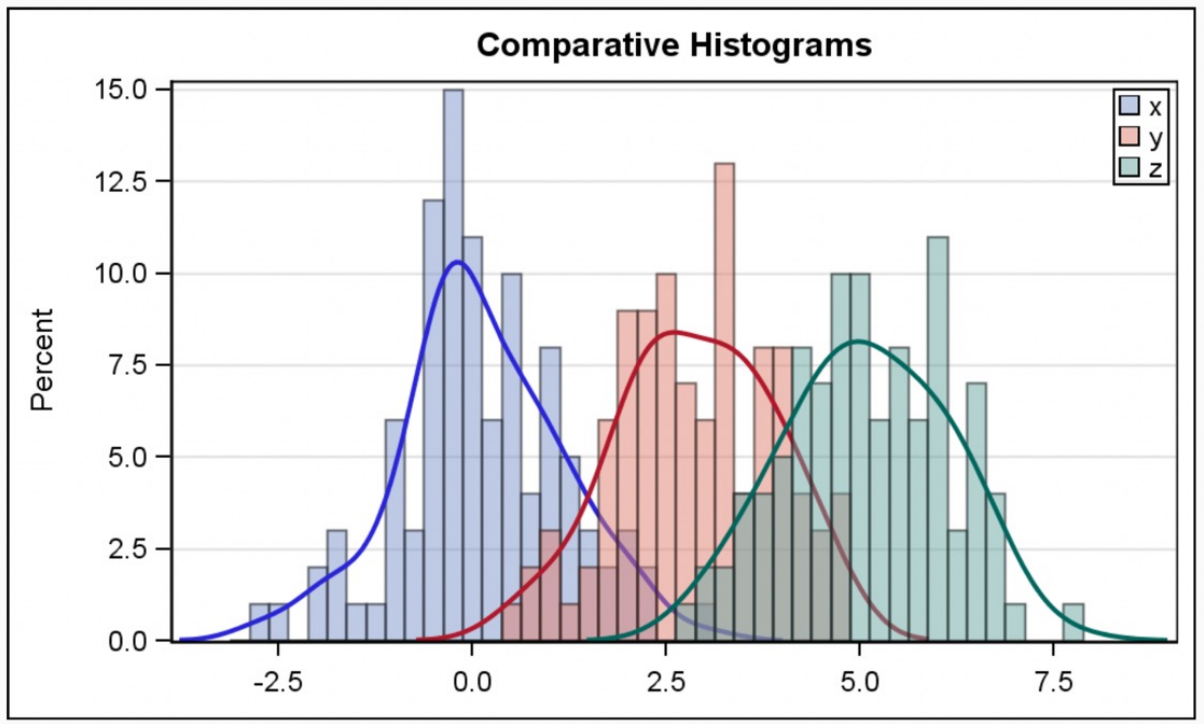
A **stacked area graph** is a way to visualize the composition of a group as it changes over time (or some other quantitative variable). This shows the relationship of a categorical variable (AgeGroup) to a quantitative variable (year).



Material provided by the Data Science course at Harvard University

Multiple Histograms

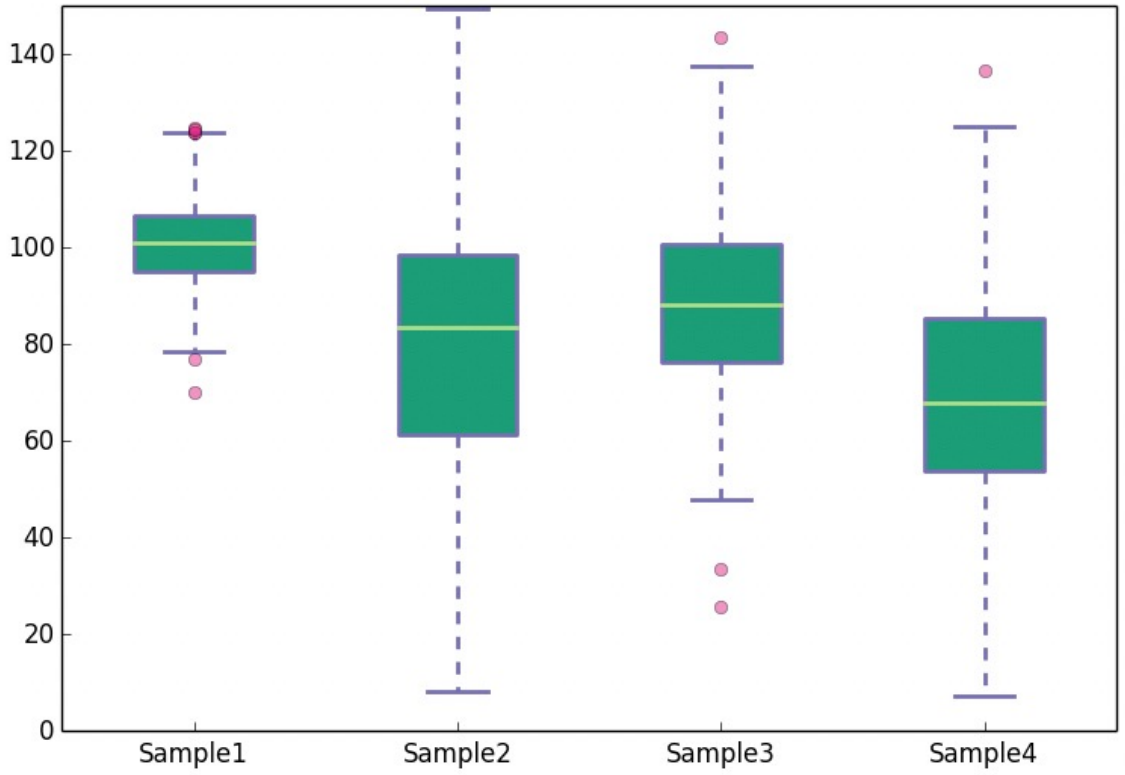
Plotting **multiple histograms (**and **kernel density estimates** of the distribution, here) on the same axes is a way to visualize how different variables compare (or how a variable differs over specific groups).



Material provided by the Data Science course at Harvard University

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.

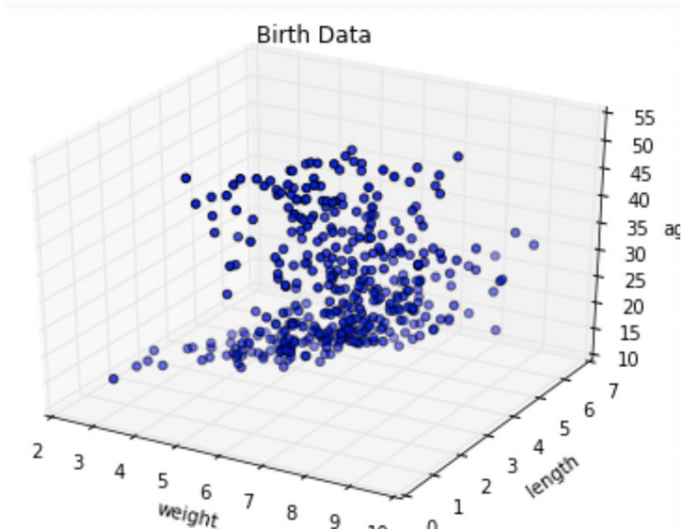


Material provided by the Data Science course at Harvard University

Not Everything is Possible!

Often your dataset seem too complex to visualize:

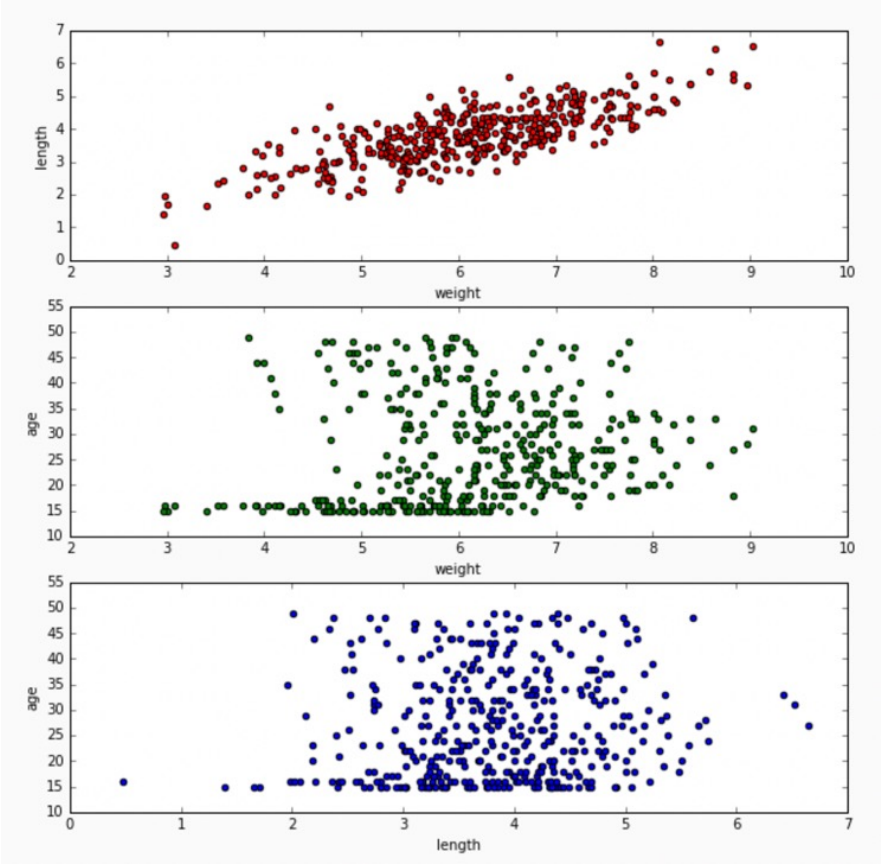
- Data is too high dimensional (how do you plot 100 variables on the same set of axes?)
- Some variables are categorical (how do you plot values like Cat or No?)



Unhelpful

Reducing Complexity

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

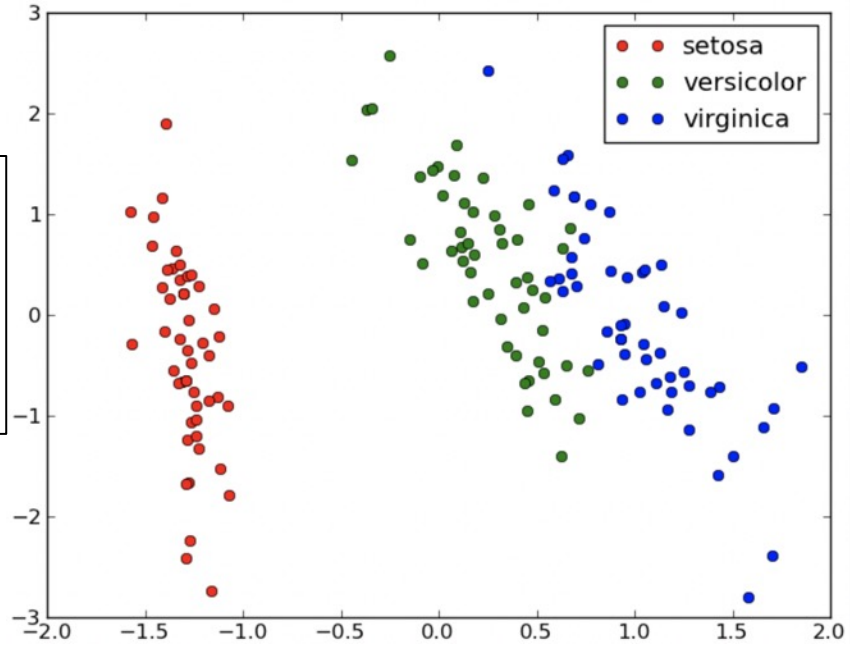


Material provided by the Data Science course at Harvard University

Reducing Complexity

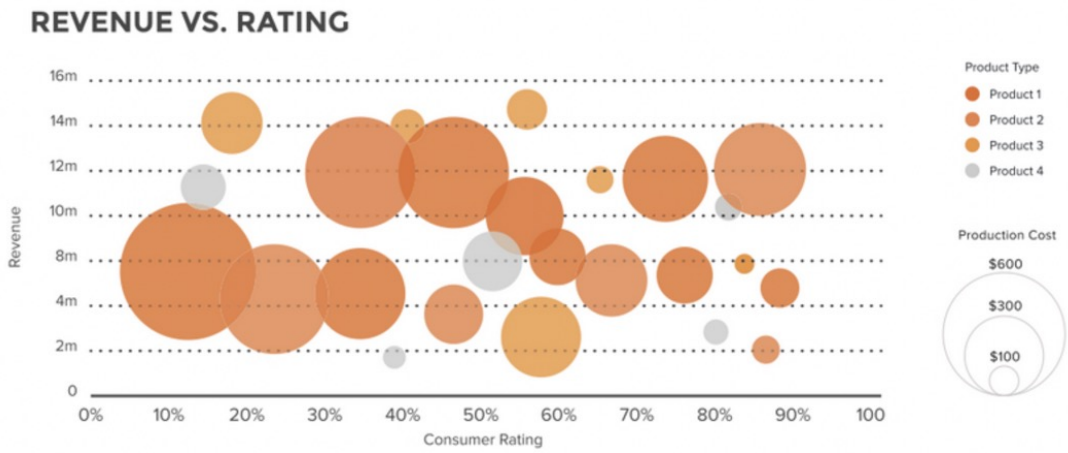
For 3D data, color coding a categorical attribute can be “effective”

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



3D can work

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

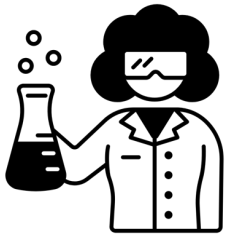


The above visualizes a set of consumer products. The variables are revenue, consumer rating, product type and product cost.

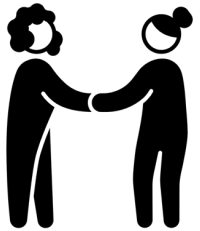
Break

CHILL

WALK



COFFEE OR TEA

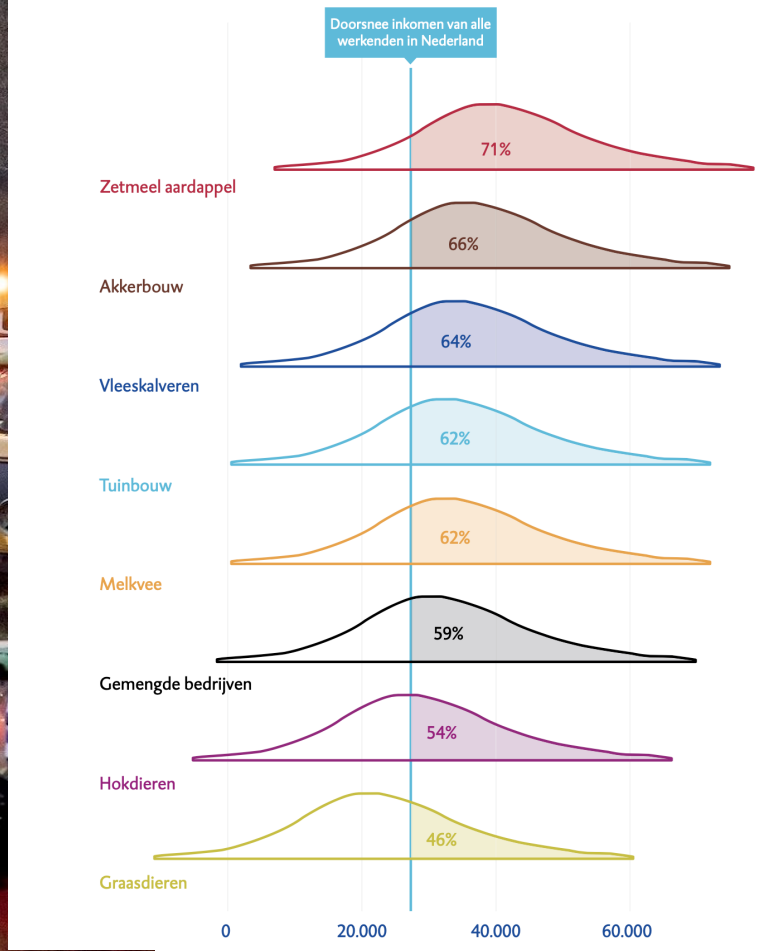


MAKE FRIENDS

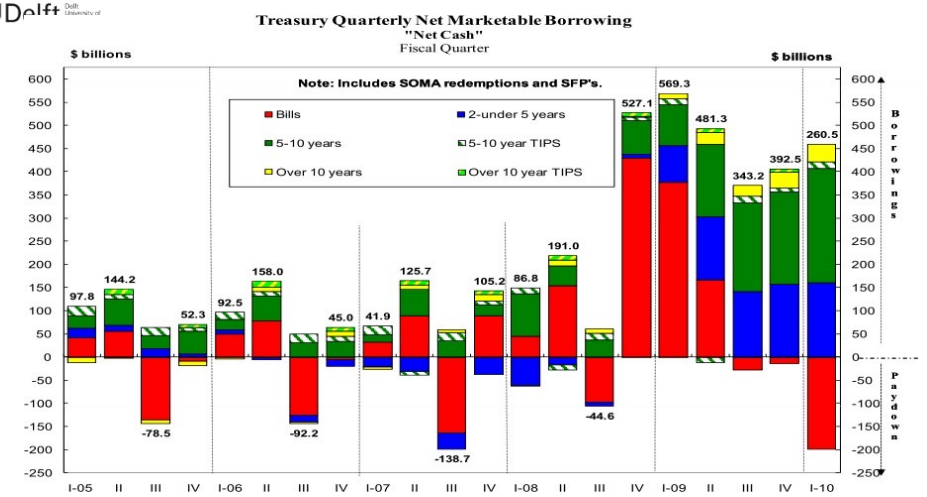
Effective Visualisation

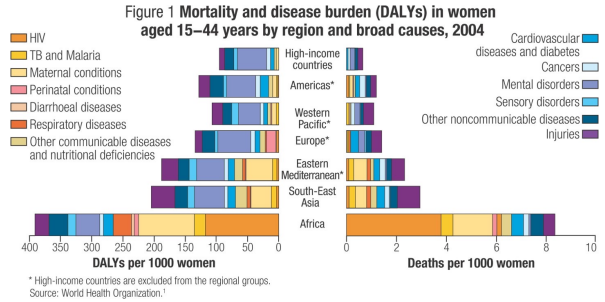
“The greatest value of a picture is when it forces us to notice what we never expected to see”

John Tukey



[Source]





Not Effective…

Sources: US Treasury and WHO reports

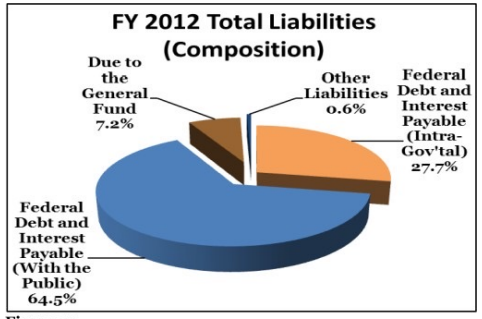
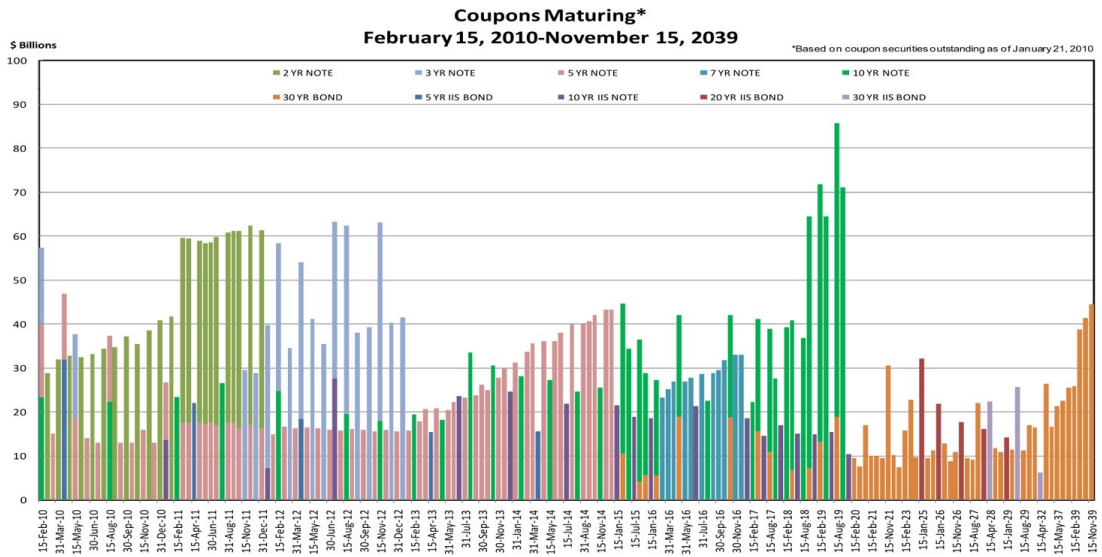


Figure 10



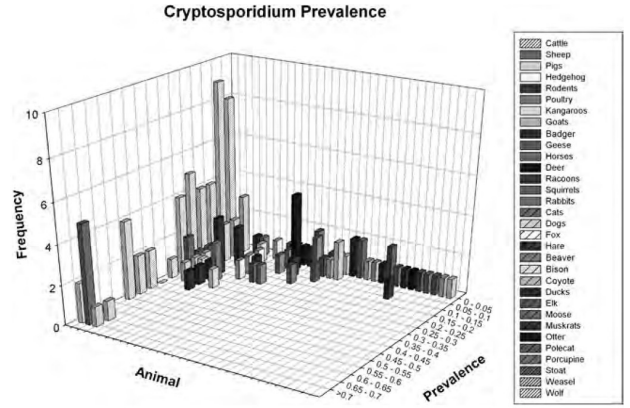


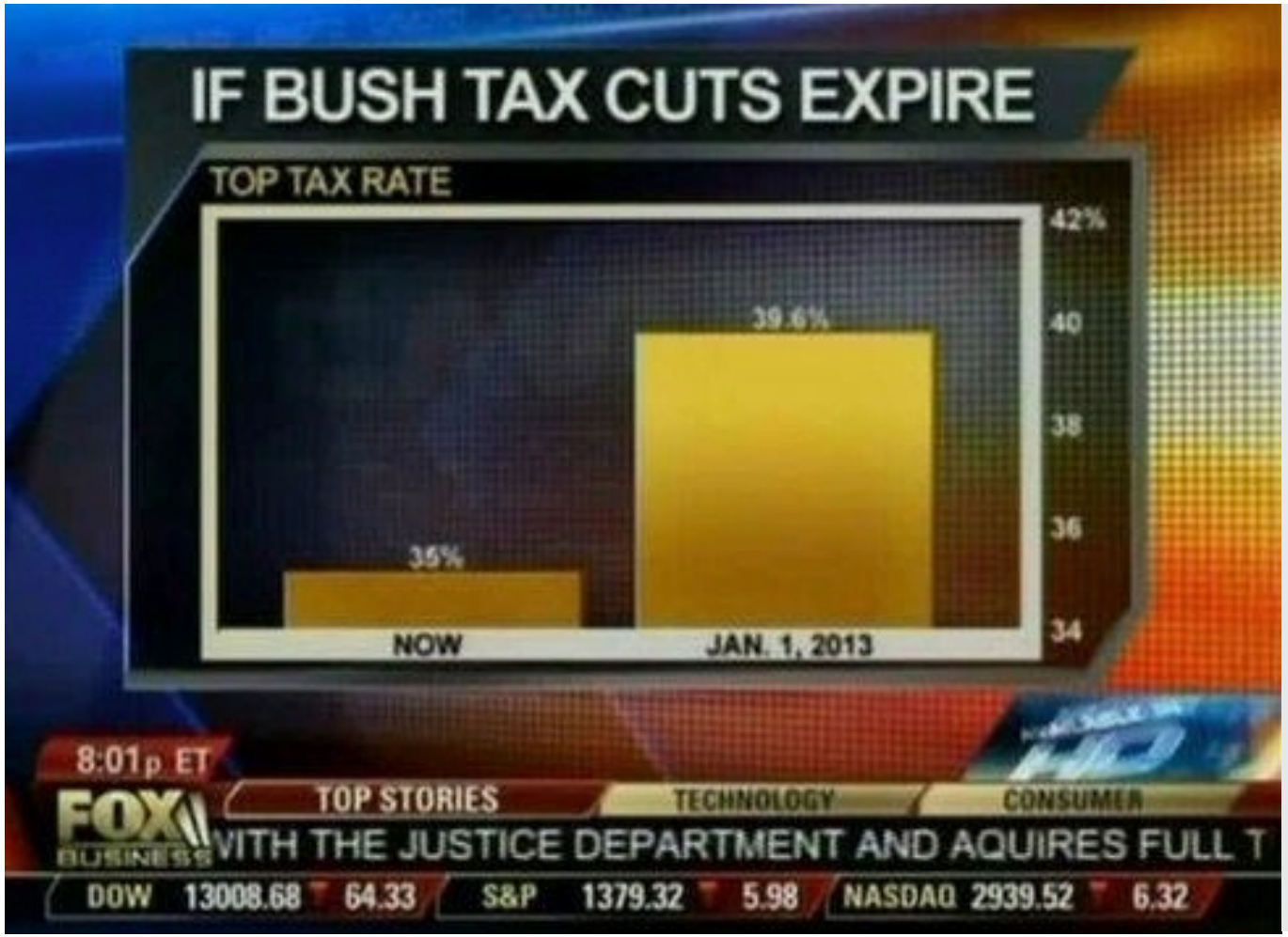
Figure 5.2 Mean prevalence rates of *Cryptosporidium* oocysts by animal species.

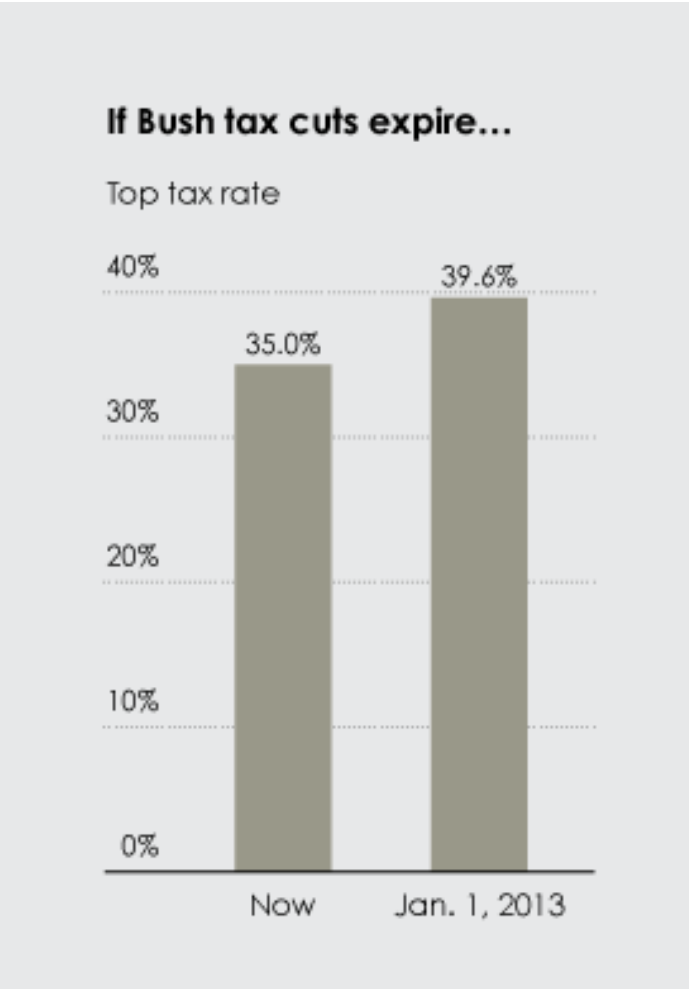
Effective EDA Visualisation

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



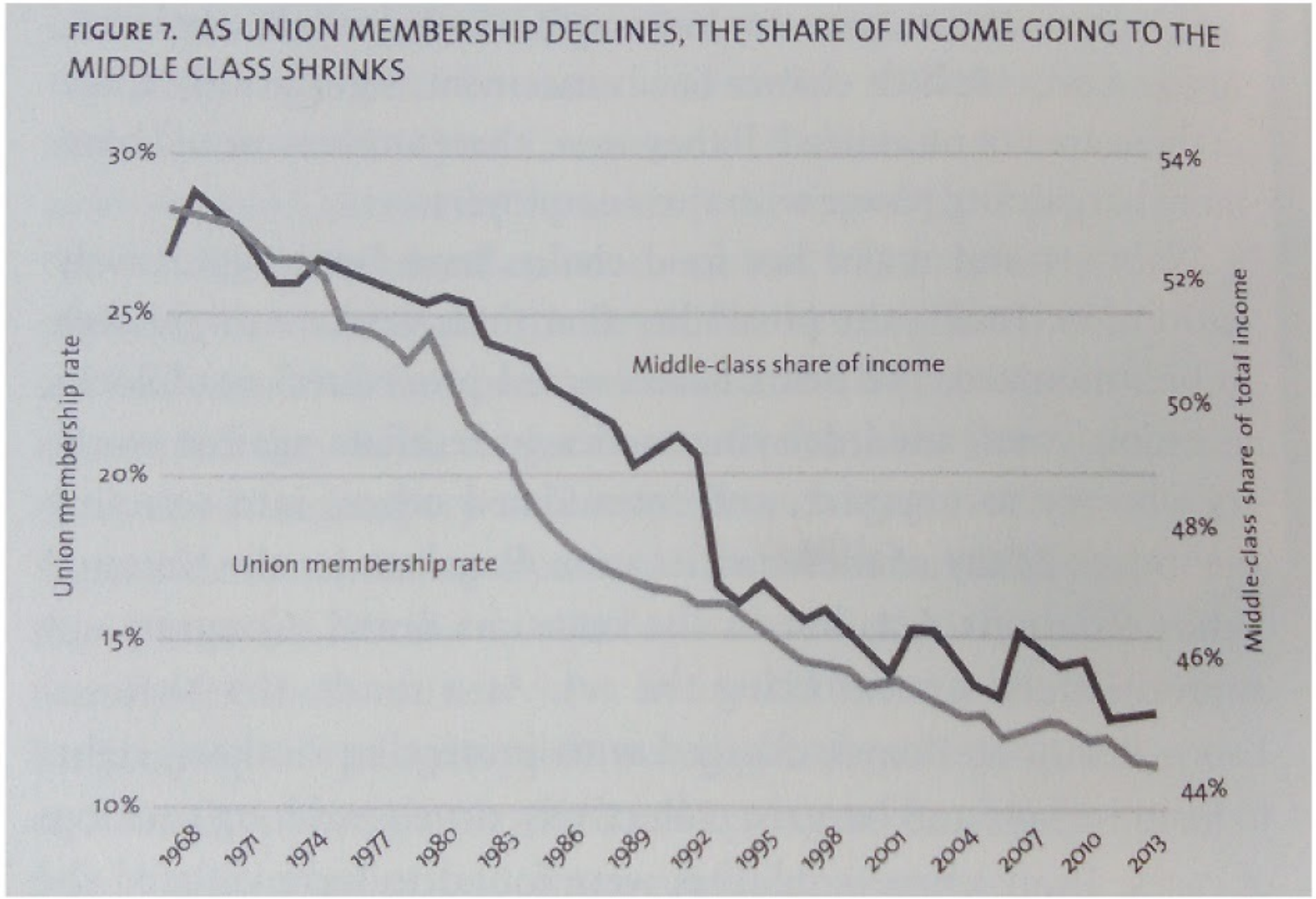


Source: Fox News

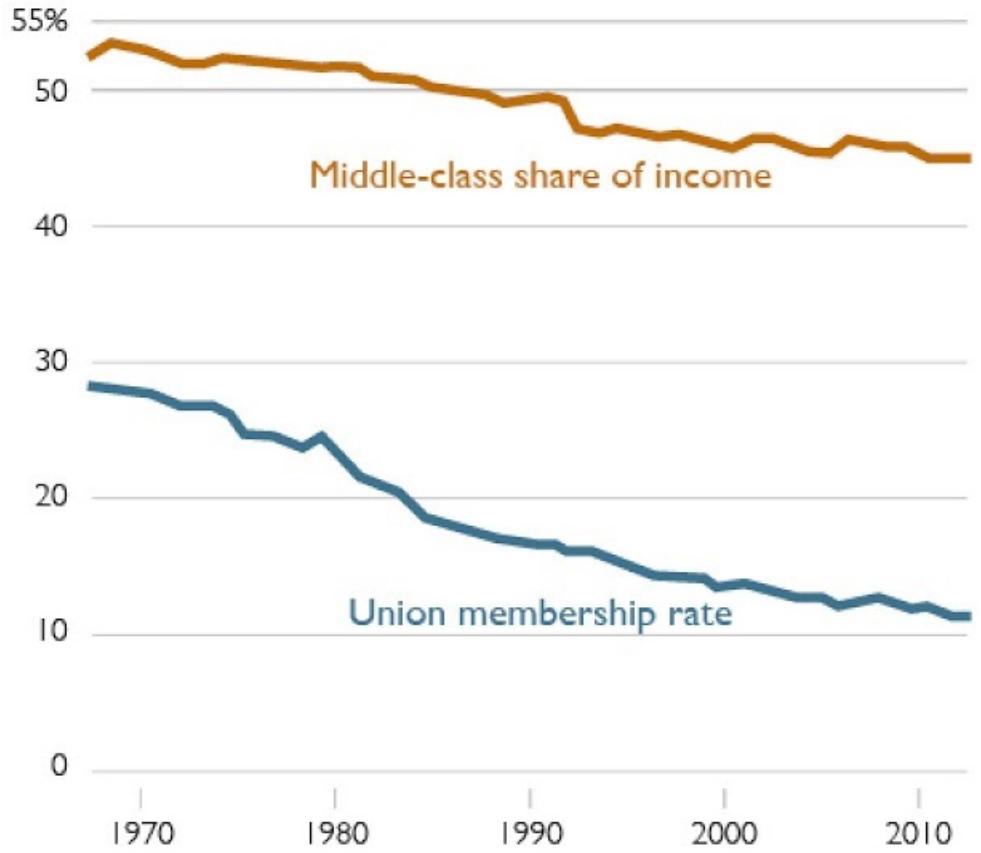
Scale Distortions

Source: Fox News

“Double the axes, double the mischief”



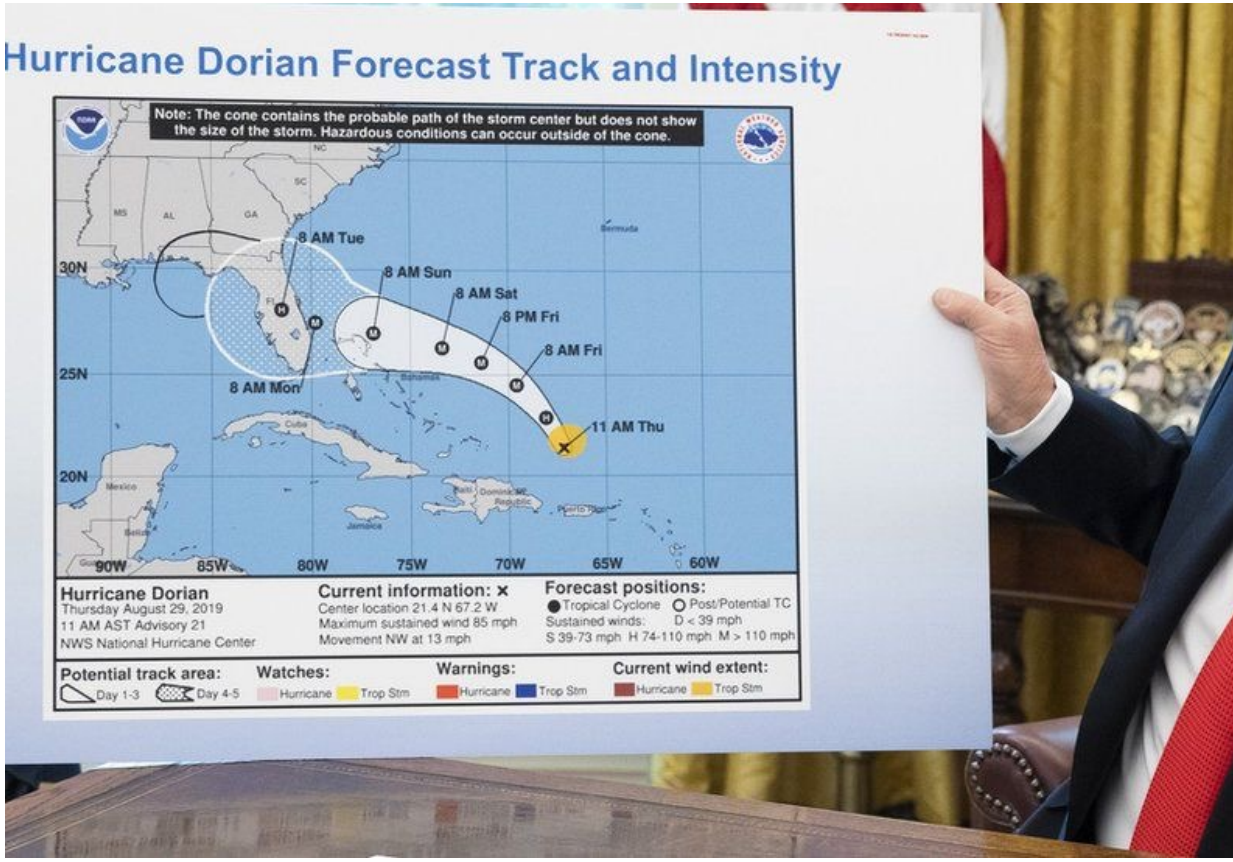
NEW VERSION



Material provided by the Data Science course at Harvard University

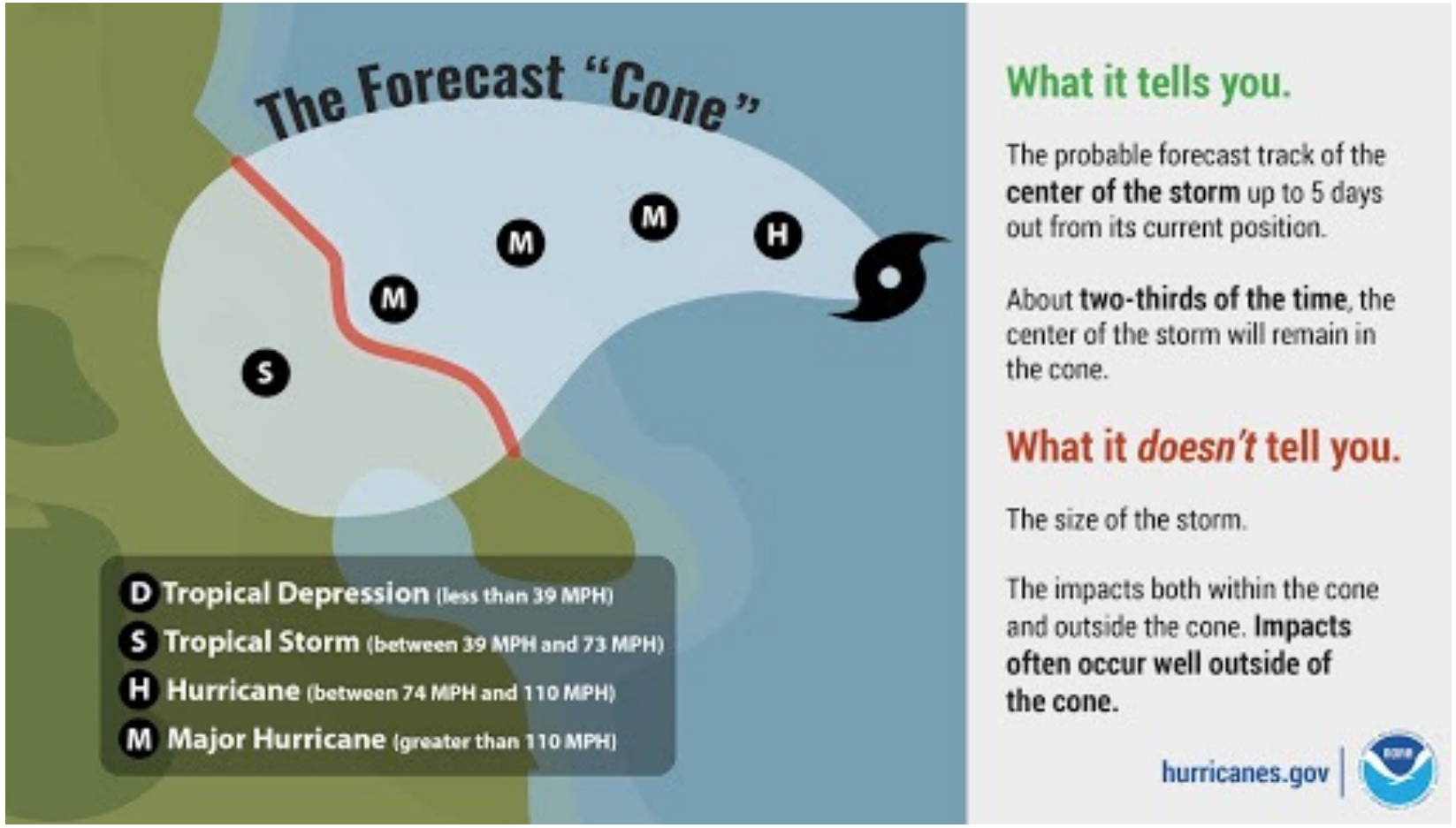
Include Uncertainty

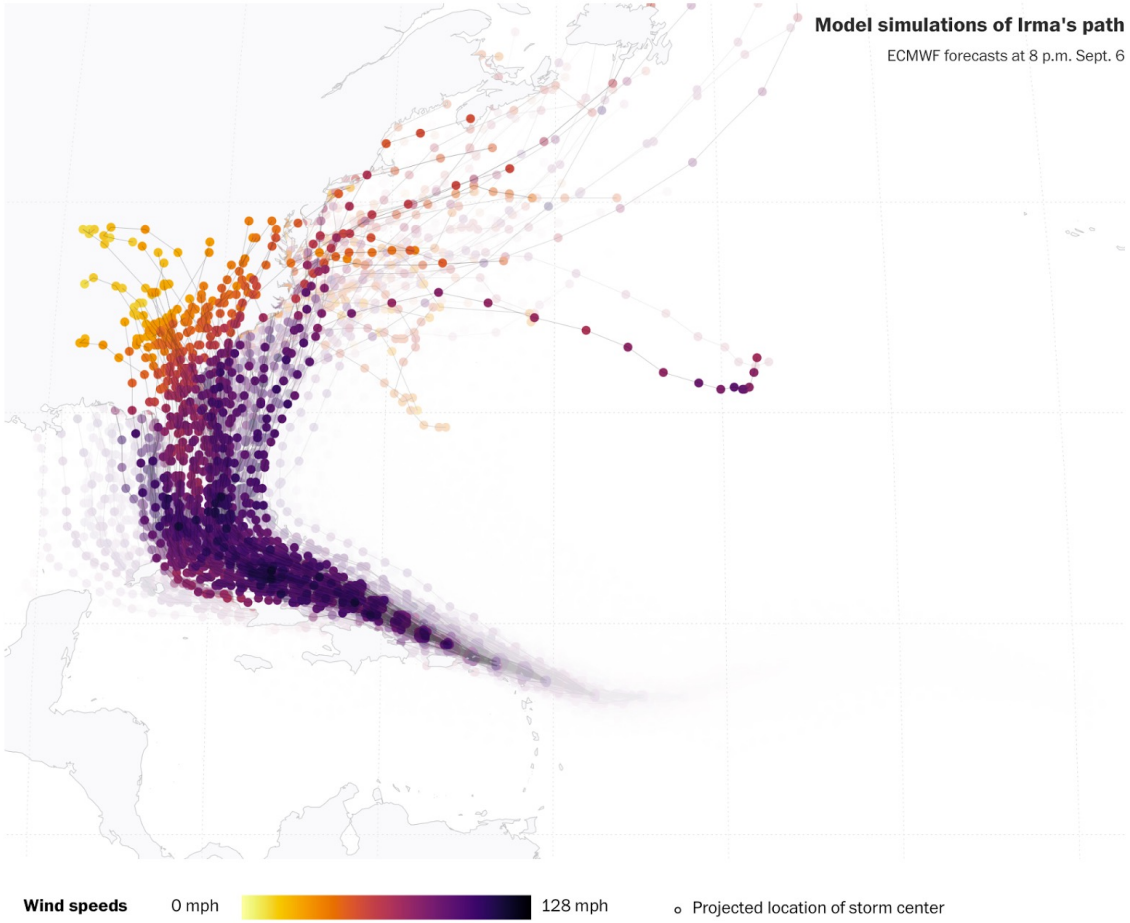
Think about Perceptions

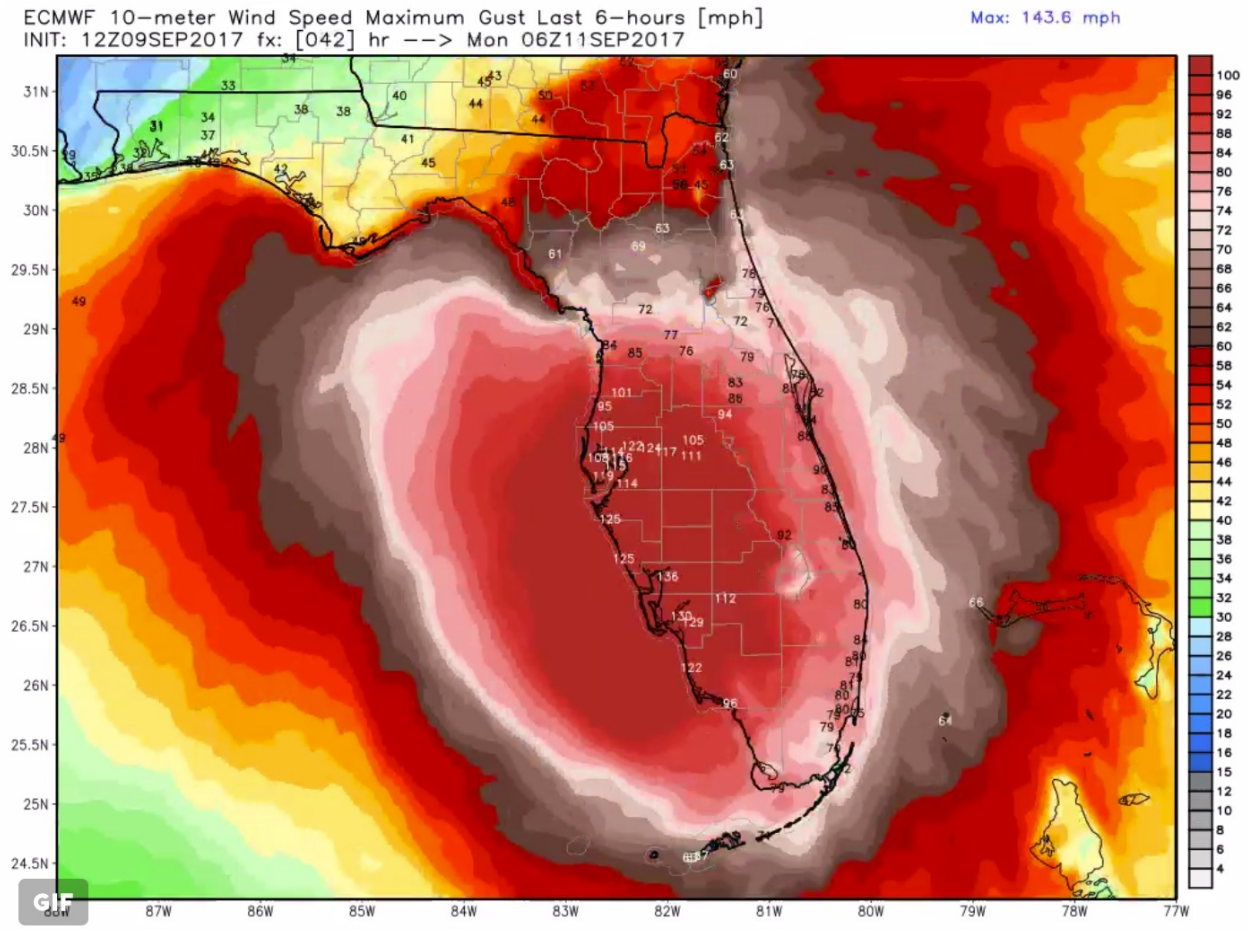


Material provided by the Data Science course at Harvard University

Think about Perceptions




Model simulations of Irma's path
ECMWF forecasts at 8 p.m. Sept. 6
Wind speeds
0 mph
128 mph
Projected location of storm center

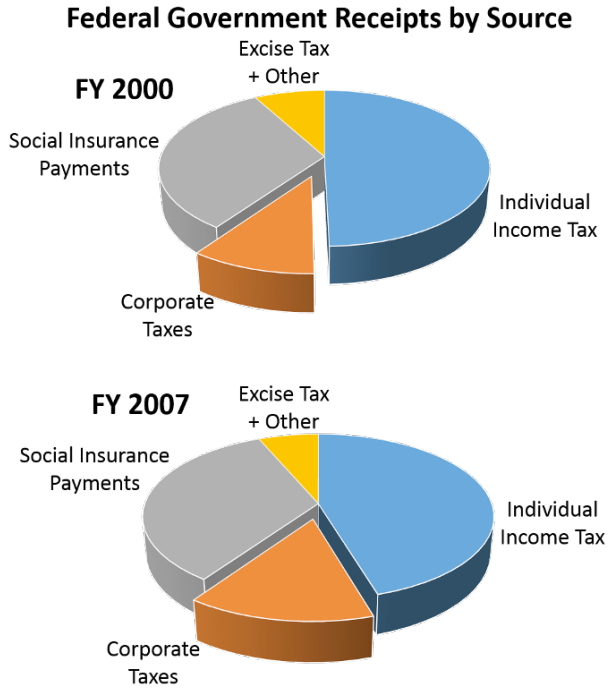

ECMWF 10-meter Wind Speed Maximum Gust Last 6-hours [mph]
INIT: 12Z09SEP2017 fx: [042] hr --> Mon 06Z11SEP2017
Max: 143.6 mph
GIF

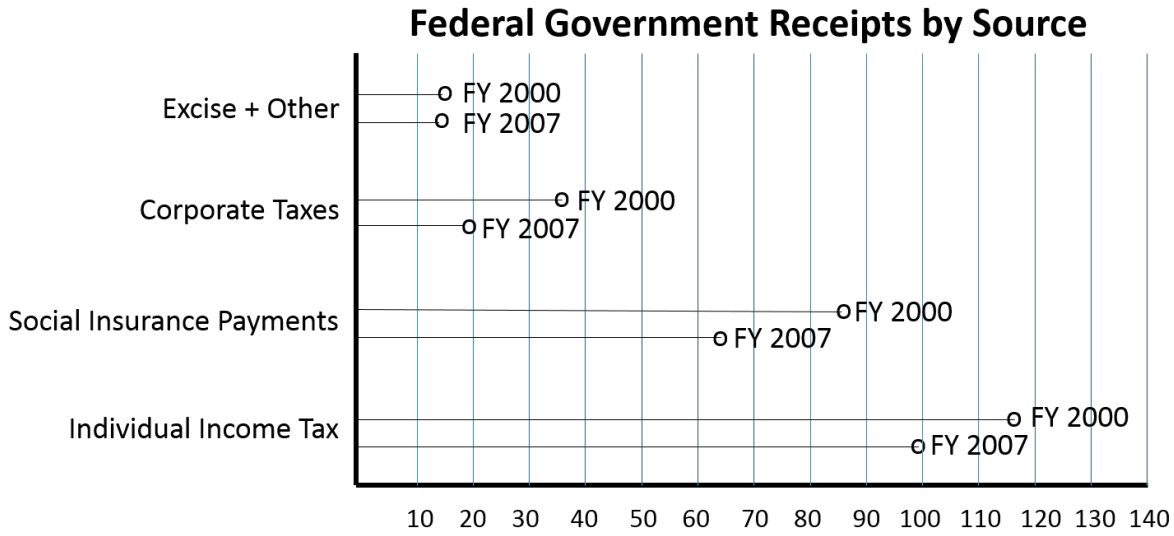
2. Keep it simple

Maximise Data-Ink Ratio – show GIF!

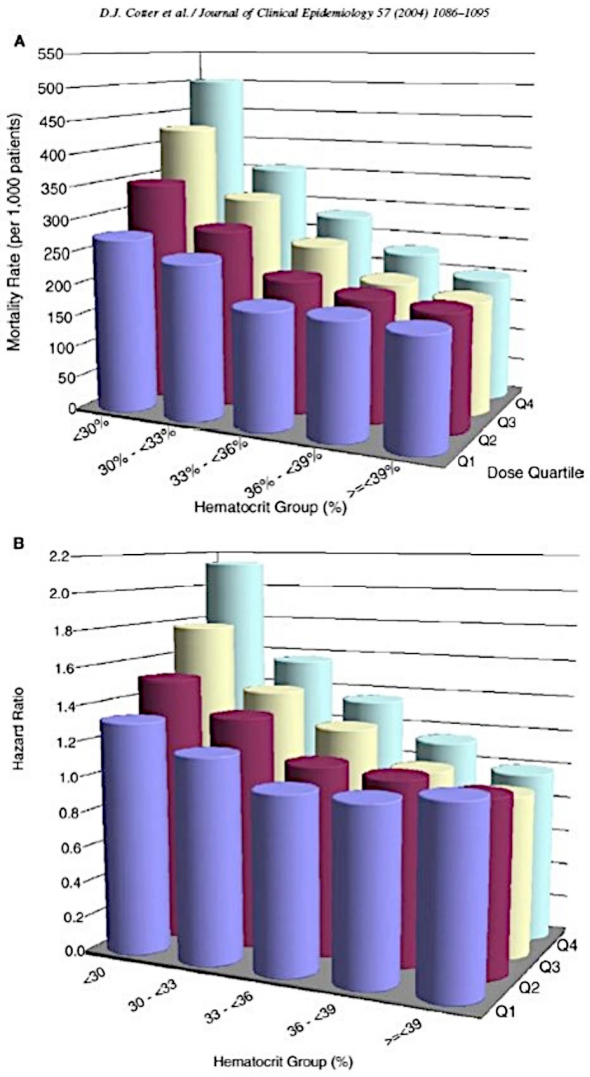
Material provided by the Data Science course at Harvard University

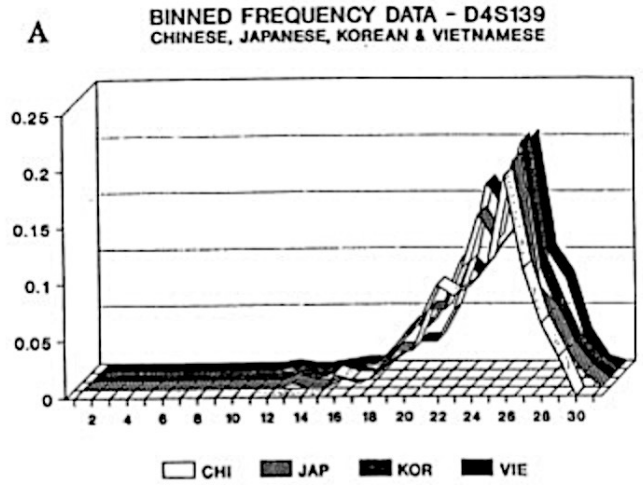
The use of Pie Charts is generally discouraged

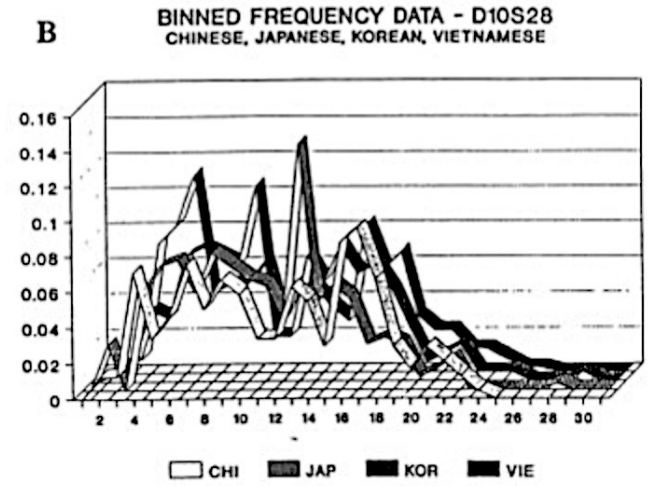




Exclude unneeded dimensions







Material provided by the Data Science course at Harvard University

Exclude unneeded dimensions

Much easier to make comparisons

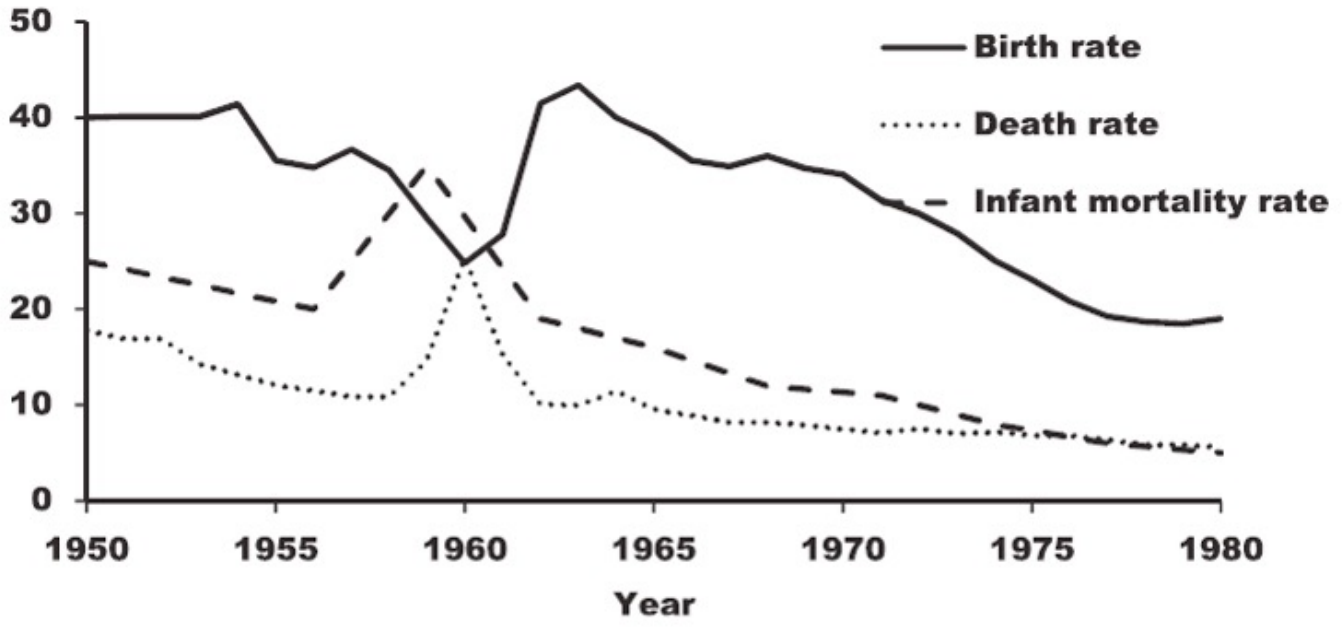
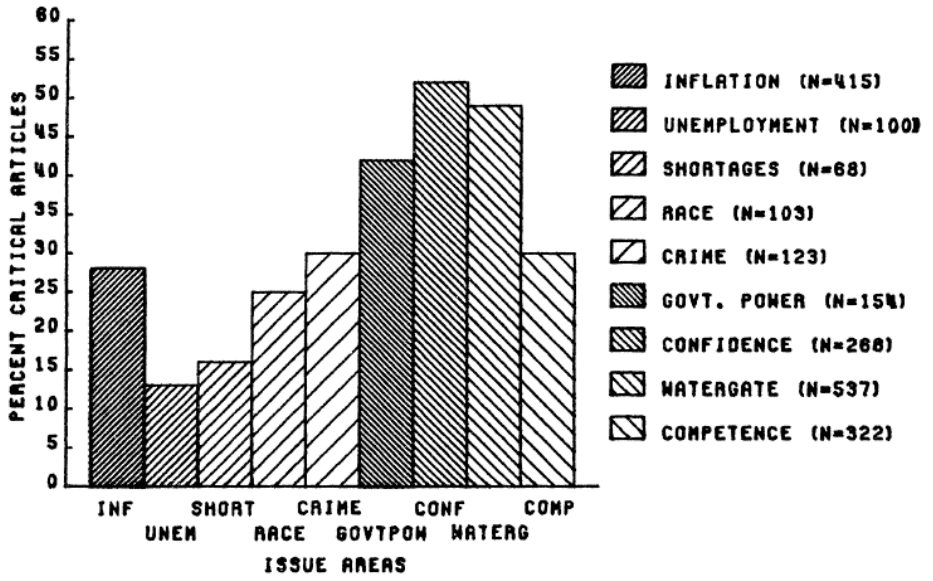


Figure 1. Fertility reduction and excess death rate and infant mortality (per thousand) during the Chinese Famine of 1959-61.
Sources: computed from the 1982 Population Census of China and the 1988 Two-Per-Thousand National Survey of Fertility and Contraception.

Omit chart junk

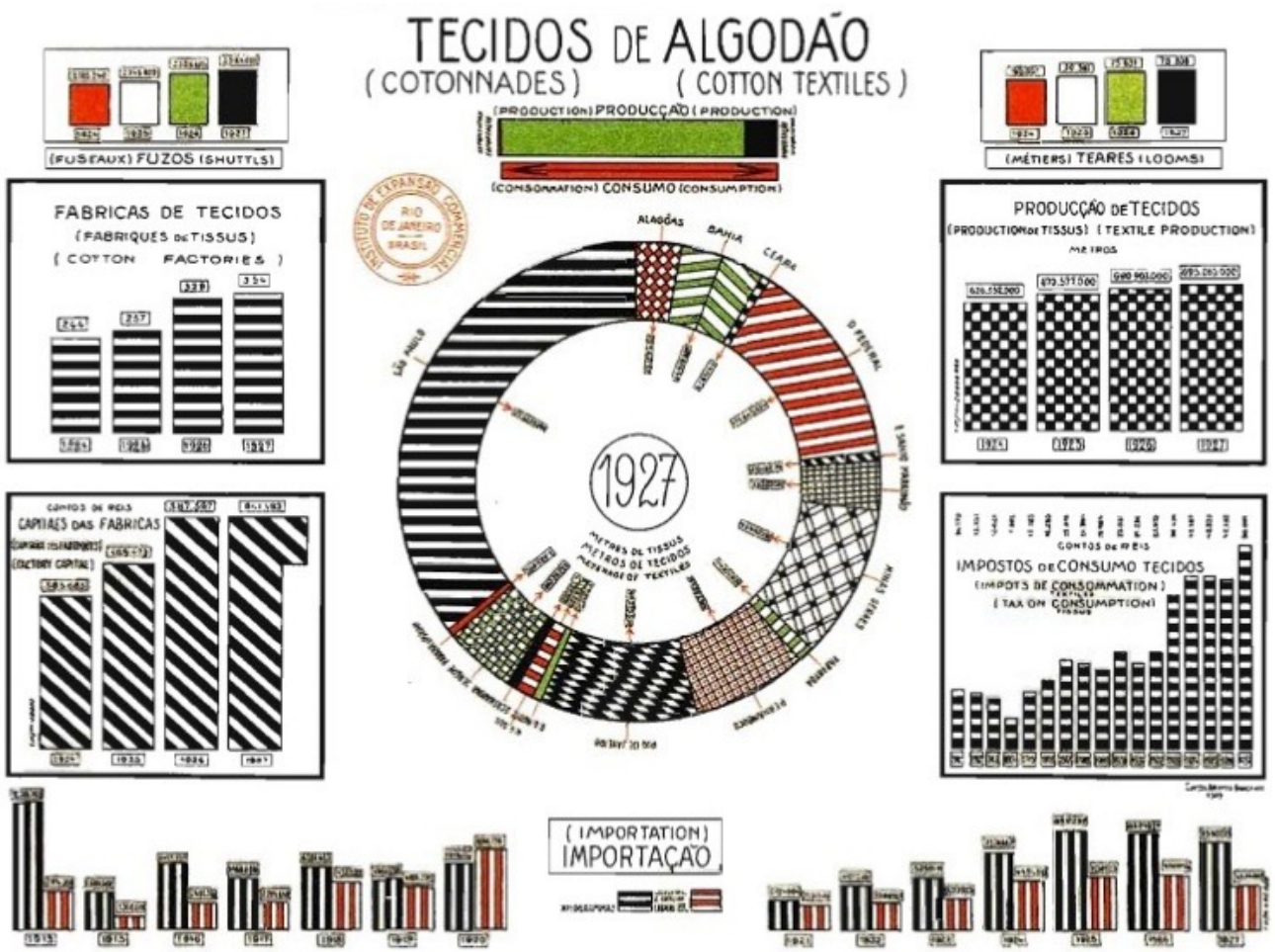


Source: Center for Political Studies Media Content Analysis Study, 1974; available through the University of Michigan, ICPSR. Not to be cited without full bibliographical reference to the present article.

- 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

Moiré vibration – visual noise, distracting

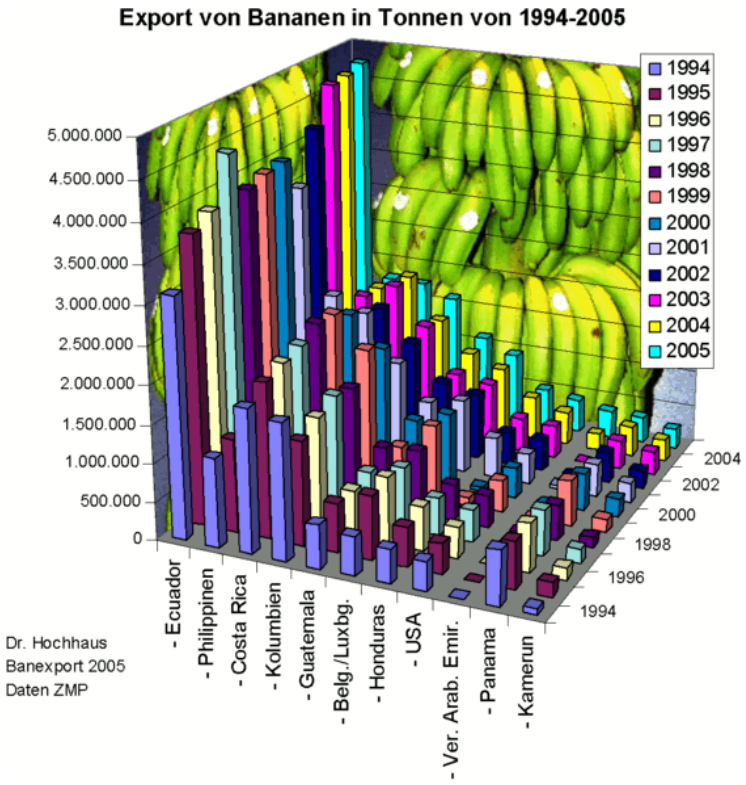


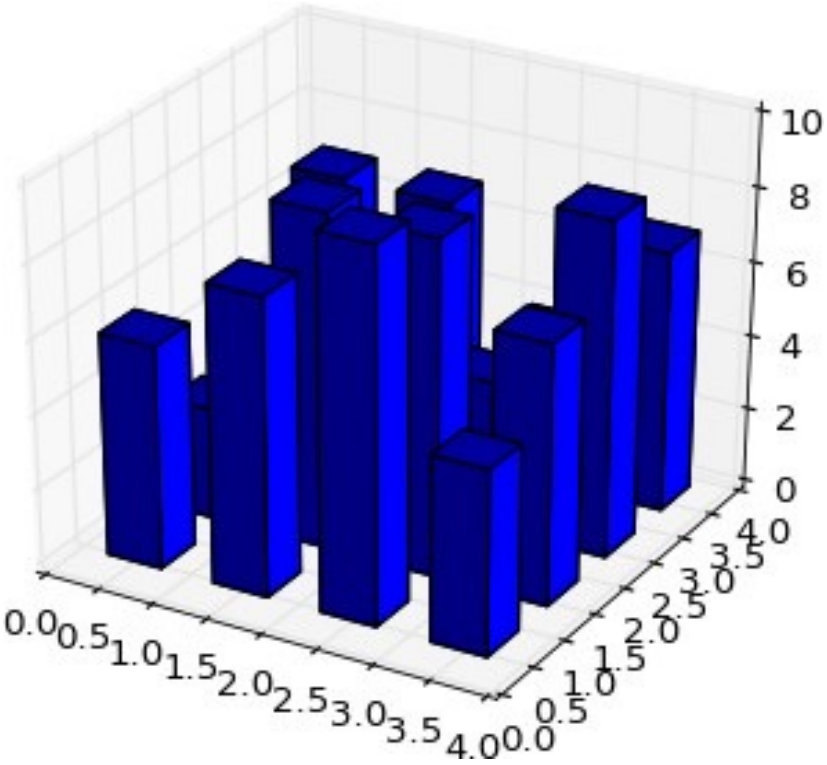
Omit chart junk

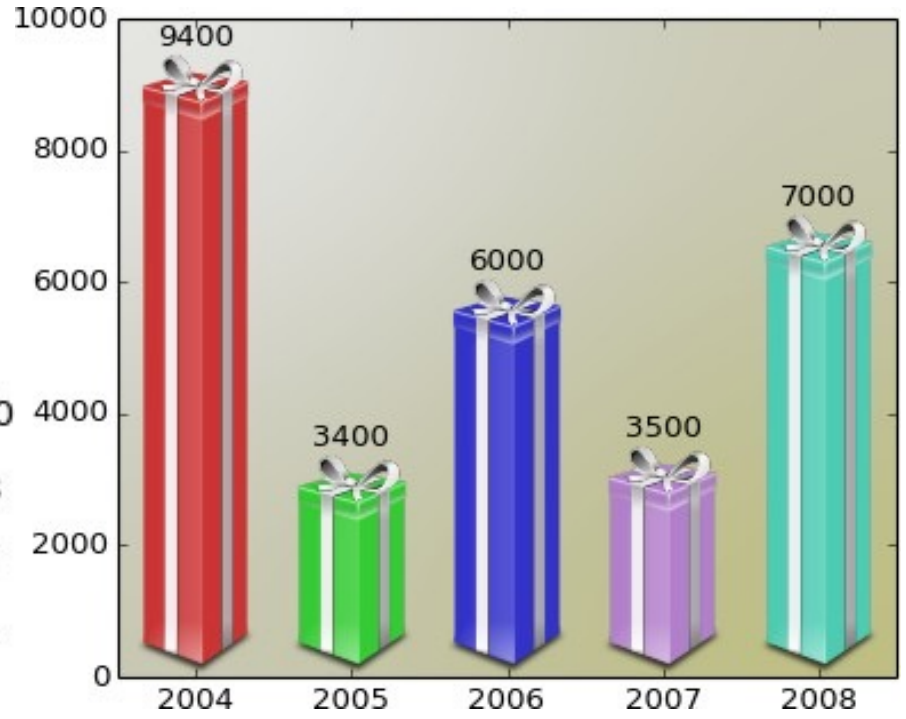
No reason to connect these counts with lines

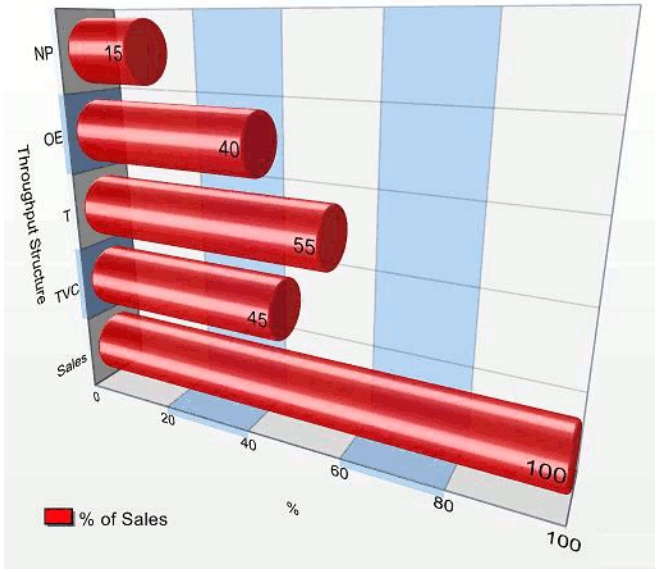
Create line graphs with Graph Maker

Don't!









Material provided by the Data Science course at Harvard University

3. Use the right display

TU Delft — Delft University of Technology

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 Bott; Liz Faunce; Graham Parrish; Billy Ehrenberg-Shannon; Paul McCallum; Martin Stabe
Inspired by the Graphic Continuum by Jon Schwabish and Severino Ribecca

ft.com/vocabulary

Deviation

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

Example FT uses
Trade surplus/deficit, climate change

Diverging bar

A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar

Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine

Splits a single value into two contrasting components (eg male/female).

Surplus/deficit filled line

The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Example FT uses
Inflation and unemployment, income and life expectancy

Scatterplot

The standard way to show the relationship between two continuous variables, each of which has its own axis.

Column + line timeline

A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot

Usually used to show how the relationship has changed over time.

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap

A good way of showing the patterns between 2 categories of data, less effective at showing fine differences in amounts.

Ranking

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

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

Ordered bar

Standard bar charts display the ranks of values much more easily when sorted into order.

Ordered column

See above.

Ordered proportional symbol

Use when there are big variations between categories and seeing fine differences between data is not so important.

Dot strip plot

Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

Slope

Perfect for showing how ranks have changed over time or vary between categories.

Lollipop

Lollipops draw more attention to the data value than standard bar/column and can also show rank and value effectively.

Bump

Effective for showing changing rankings across multiple dates. For large datasets, consider grouping lines using colour.

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population (age/sex) distribution, revealing inequality

Histogram

The standard way to show a statistical distribution – keep the gaps between columns small to highlight the 'shape' of the data.

Dot plot

A simple way of showing the change or range (min/max) of data across multiple categories.

Dot strip plot

Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Barcode plot

Like dot strip plots, good for displaying all the data in a table, they work best when highlighting individual values.

Boxplot

Summarise multiple distributions by showing the median (centre) and range of the data

Violin plot

Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple average).

Population pyramid

A standard way for showing the age and sex breakdown of a population distribution; effectively, back to back histograms.

Cumulative curve

A good way of showing how unequal a distribution is: y axis is always cumulative frequency, x axis is always a measure.

Frequency polygons

For displaying multiple distributions of data. Like a regular line chart, best limited to a maximum of 3 or 4 datasets.

Beeswarm

Use to emphasise individual points in a distribution. Points can be sized to an additional variable. Best with medium-sized datasets

Change over Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series 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

Line

The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column

Columns work well for showing change over time – but usually best with only one series of data at a time.

Column + line timeline

A good way of showing the relationship over time between an amount (columns) and a rate (line).

Slope

Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story.

Area chart

Use with care – these are good at showing changes to total, but seeing change in components can be very difficult.

Candlestick

Usually focused on day-to-day activity, these charts show opening/closing and high/low points of each day.

Fan chart (projections)

Use to show the uncertainty in future projections – usually this grows the further forward to projection.

Connected scatterplot

A good way of showing changing data for two variables whenever there is a relatively clear pattern of progression.

Calendar heatmap

A great way of showing temporal patterns (daily, weekly, monthly) – at the expense of showing precision in quantity.

Priestley timeline

Great when date and duration are key elements of the story in the data.

Circle timeline

Good for showing discrete values of varying size across multiple categories (eg earthquakes by continent).

Vertical timeline

Presents time on the Y axis. Good for displaying detailed time series that work especially well when scrolling on mobile.

Seismogram

Another alternative to the circle timeline for showing series where there are big variations in the data.

Streamgraph

A type of area chart; use when seeing changes in proportions over time is more important than individual values

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Example FT uses
Commodity production, market capitalisation, volumes in general

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Bar

See above. Good when the data are not time series and labels have the same length – but usually best with category names.

Paired column

As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar

See above.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Isotype (pictogram)

Excellent solution in some instances – use only with whole numbers (do not slice off an arm to represent a decimal).

Lollipop

Lollipop charts draw more attention to the data value than standard bar/column – does not have to start at zero (but preferable).

Radar

A space-efficient way of showing value of multiple variables– but make sure they are organised in a way that makes sense to reader.

Parallel coordinates

An alternative to radar charts – again, the arrangement of the variables is important. Usually benefits from highlighting values.

Bullet

Good for showing a measurement against the context of a target or performance range.

Grouped symbol

An alternative to bar/column charts when being able to count data or highlight individual elements is useful.

Part-to-whole

Show how a single entity can be broken down into its component elements. 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

Stacked column/bar

A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Pie

A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut

Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg total).

Treemap

Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments.

Voronoi

A way of turning points into areas – any point within each area is closer to the central point than any other centroid.

Arc

A hemicycle, often used for visualising parliamentary composition by number of seats.

Gridplot

Good for showing % information, they work best when used on whole numbers and work well in small multiple layout form.

Venn

Generally only used for schematic representation.

Waterfall

Can be useful for showing part-to-whole relationships where some of the components are negative.

Spatial

Aside from locator maps only used when precise locations or geographical patterns in data are more important to the reader than anything else.

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

Basic choropleth (rate/ratio)

The standard approach for putting data on a map – should always be rates rather than totals and use a sensible base geography.

Proportional symbol (count/magnitude)

Use for totals rather than rates – be wary that small differences in data will be hard to see.

Flow map

For showing unambiguous movement across a map.

Contour map

For showing areas of equal value on a map. Can use deviation colour schemes for showing +/- values

Equalised cartogram

Converting each unit on a map to a regular and equally-sized shape – good for representing voting regions with equal value.

Scaled cartogram (value)

Stretching and shrinking a map so that each area is sized according to a particular value.

Dot density

Used to show the location of individual events/locations – make sure to annotate any patterns the reader should see.

Heat map

Grid-based data values mapped with an intensity colour scale. As choropleth map – but not snapped to an admin/political unit.

Flow

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses
Movement of funds, trade, migrants, lawsuits, information; relationship graphs.

Sankey

Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

Waterfall

Designed to show the sequencing of data through a flow process, typically budgets. Can include +/- components.

Chord

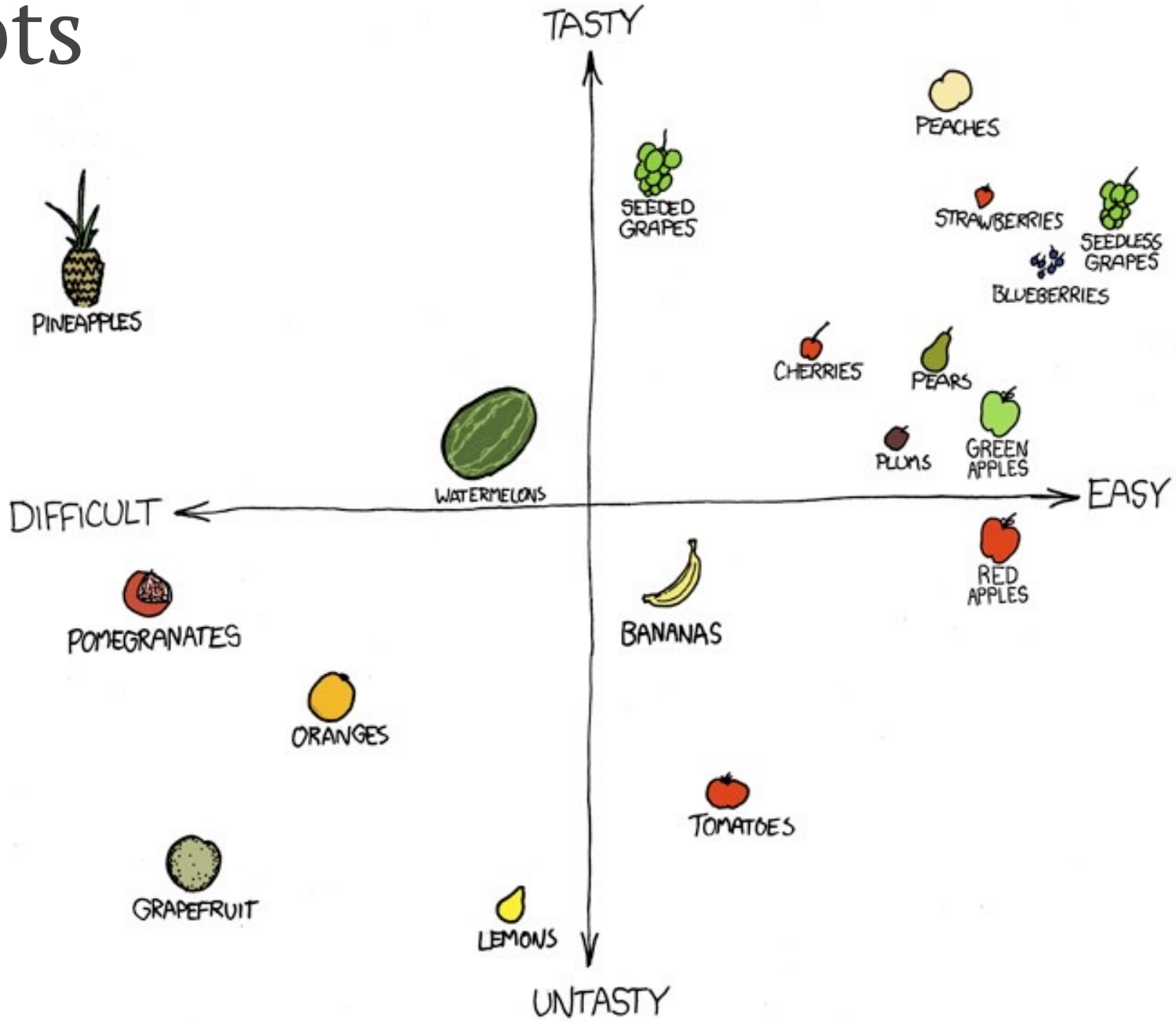
A complex but powerful diagram which can illustrate 2-way flows (and net winner) in a matrix.

Network

Used for showing the strength and inter-connectedness of relationships of varying types.

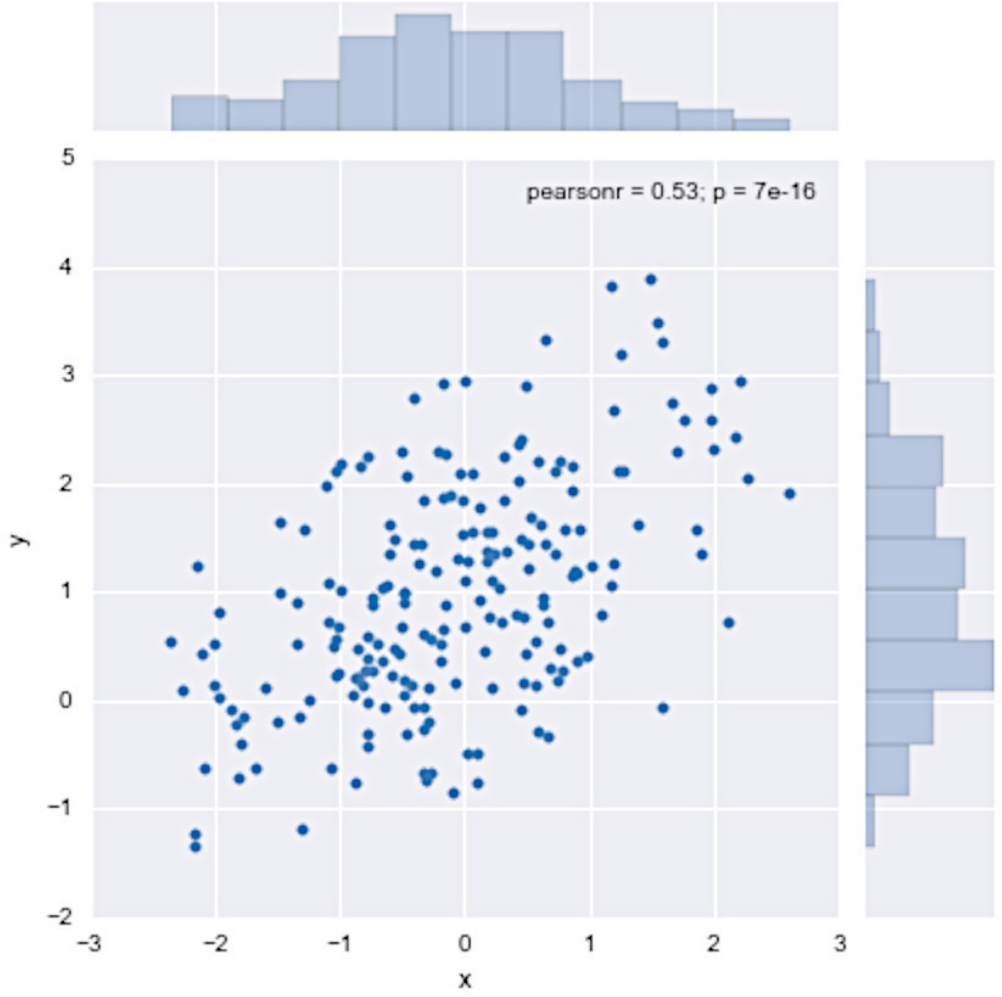
Correlations

Scatterplots

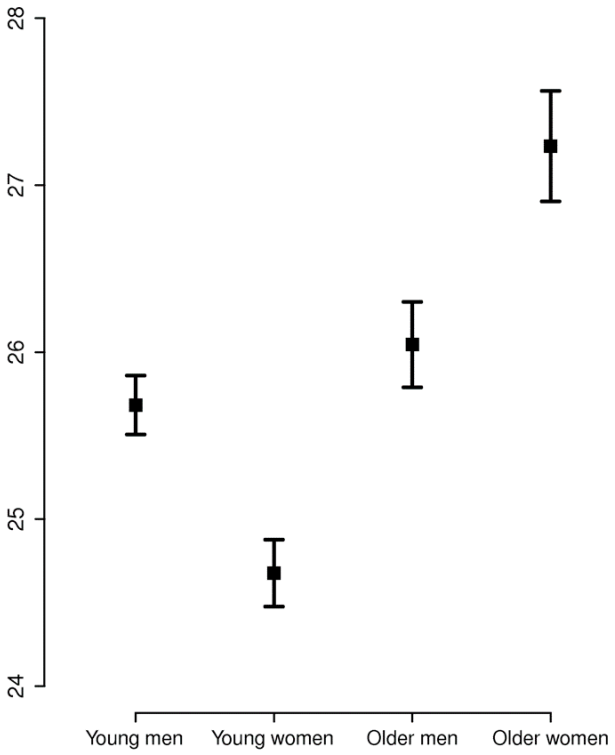


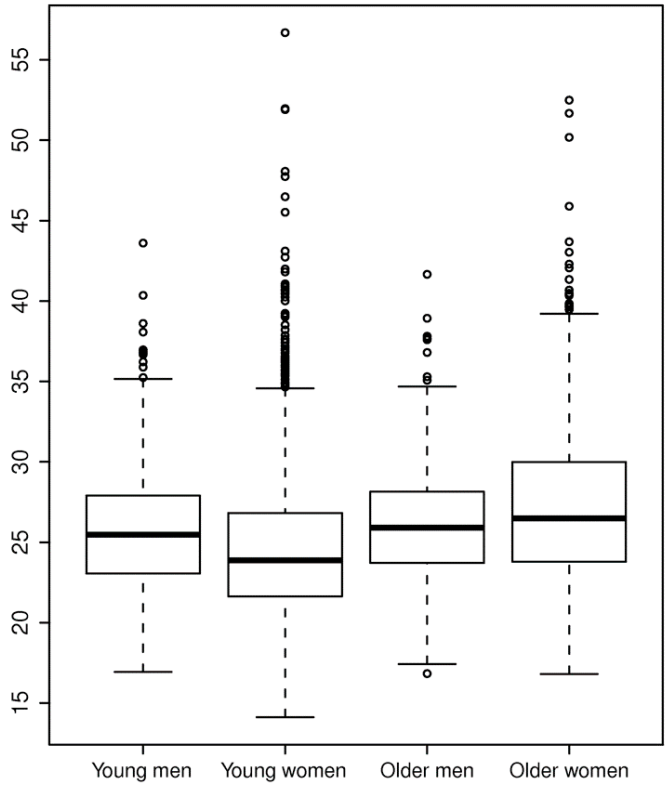
Material provided by the Data Science course at Harvard University

Make efficient use of space



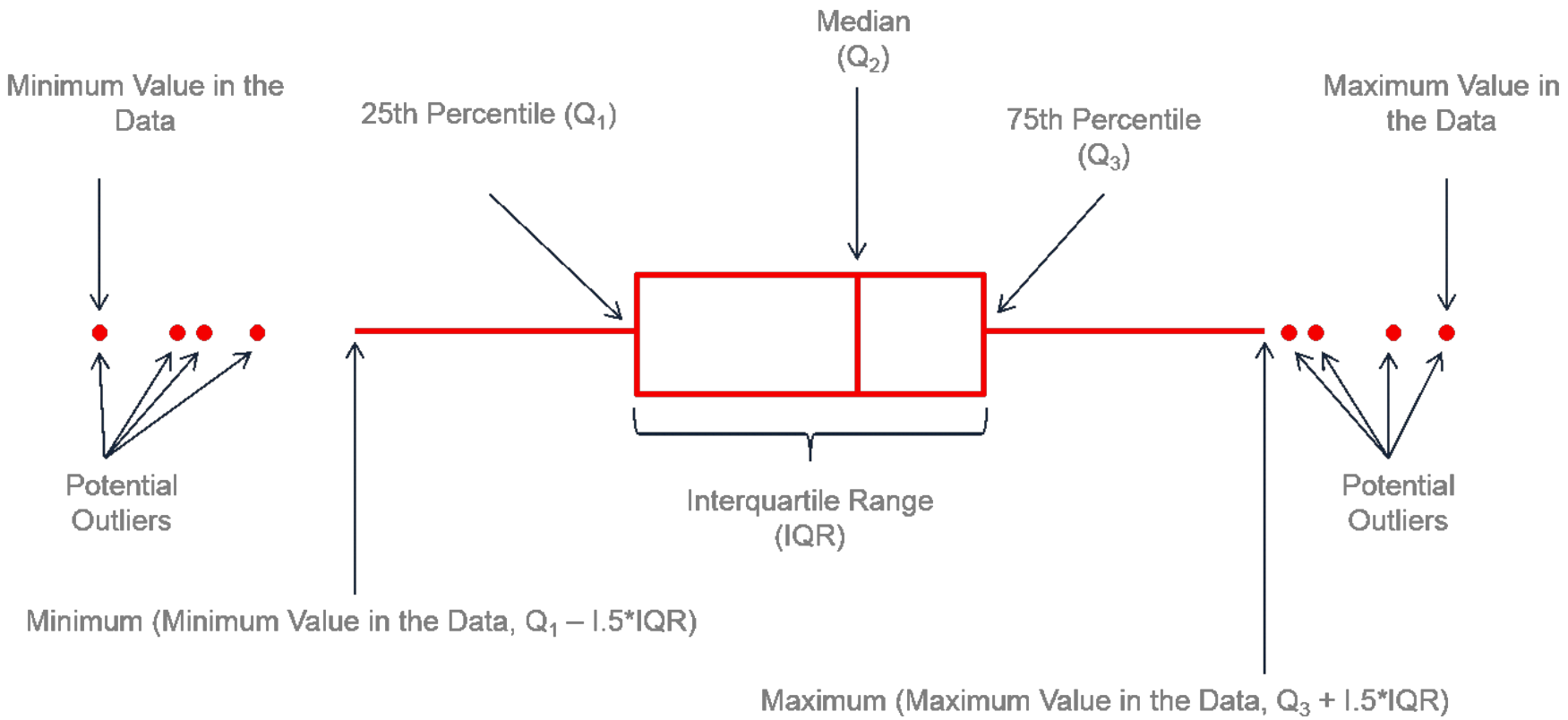
Make efficient use of space





- 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.

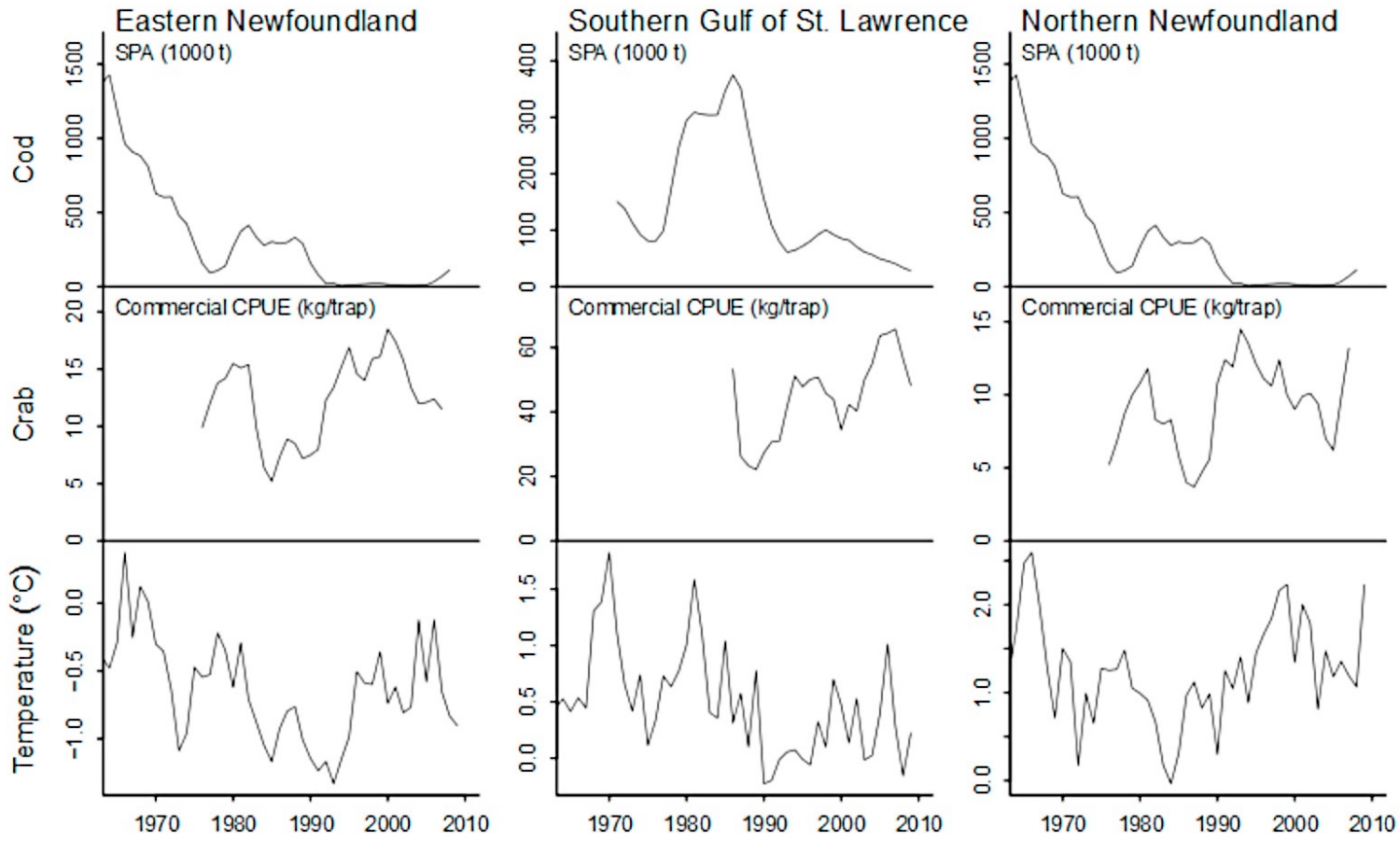
Make efficient use of space



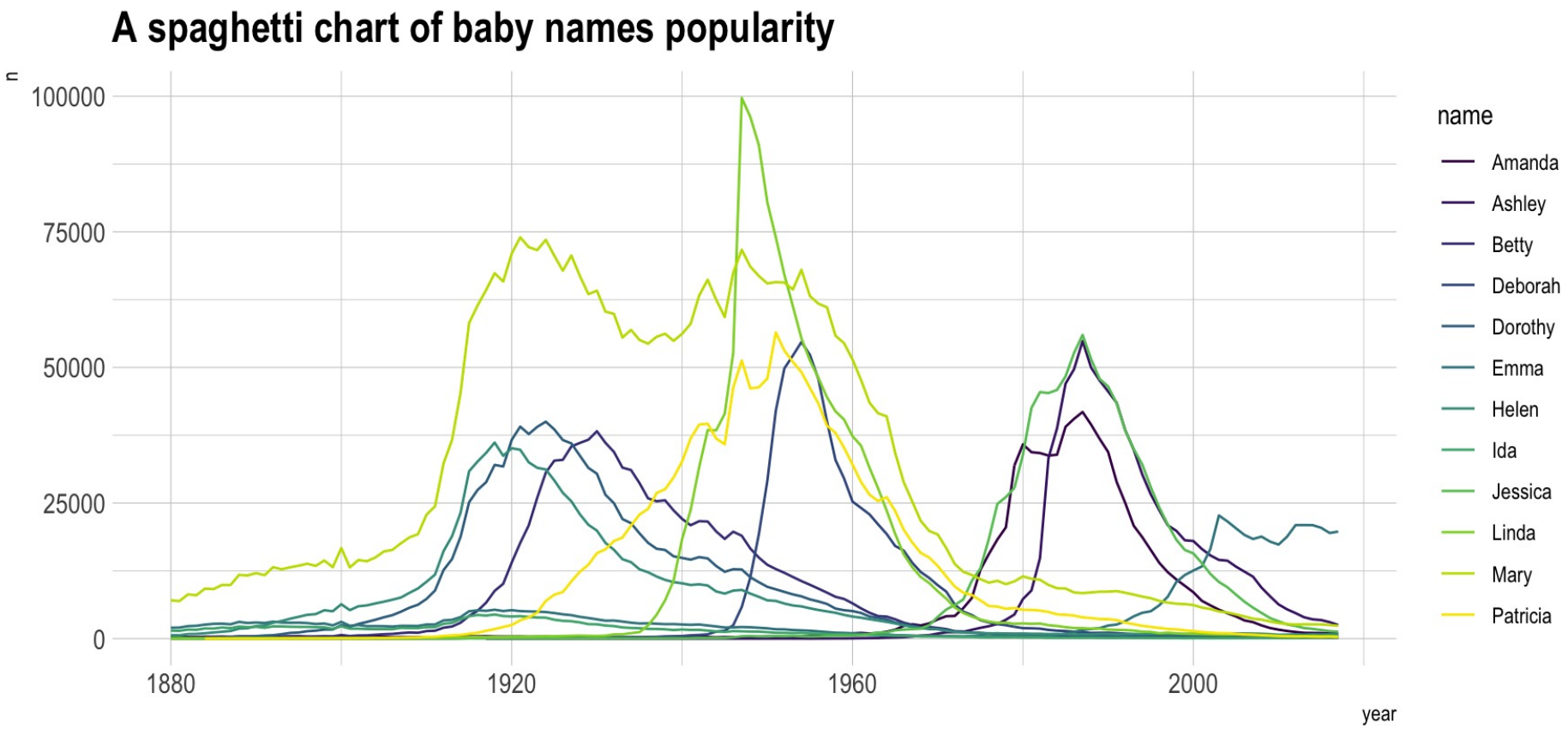
Bar charts are not appropriate for indicating means ± SEs: they add ink without conveying any additional information

Facilitate Comparison

[Y] axes are different

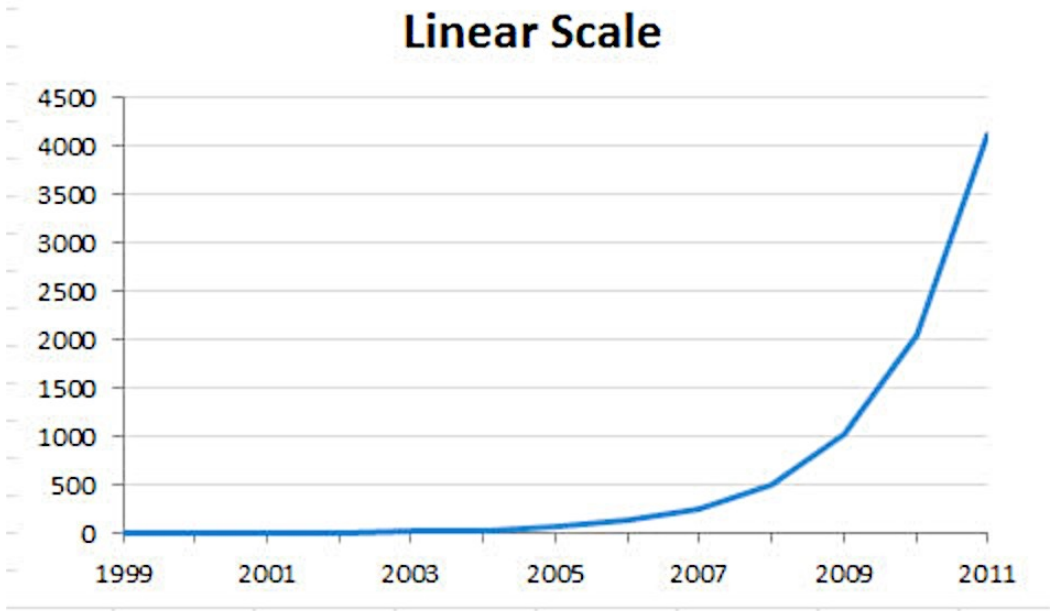


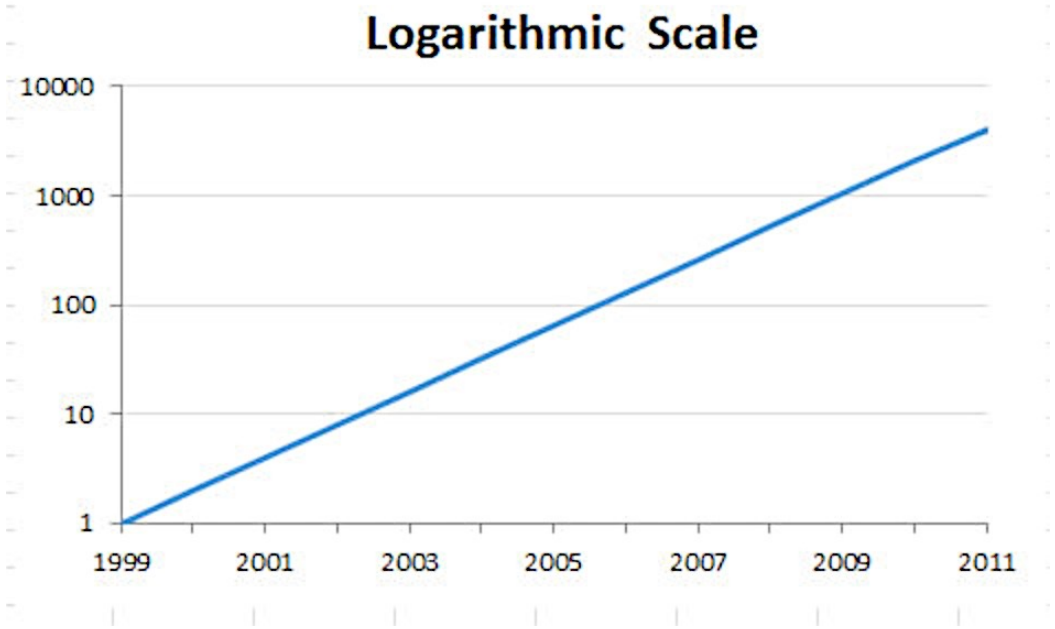
Facilitate Comparison



Material provided by the Data Science course at Harvard University

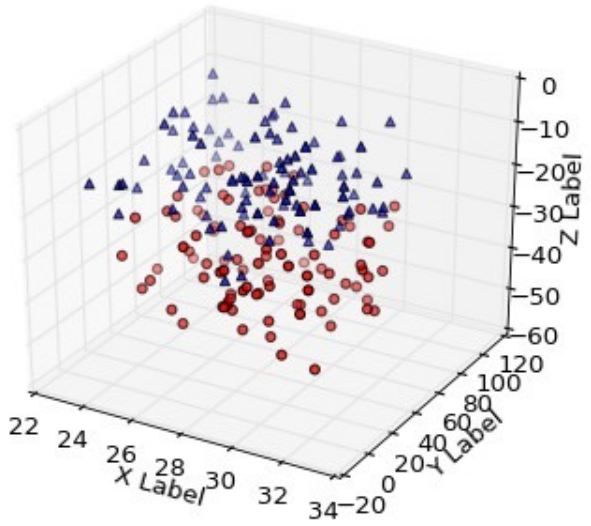
Facilitate Comparison



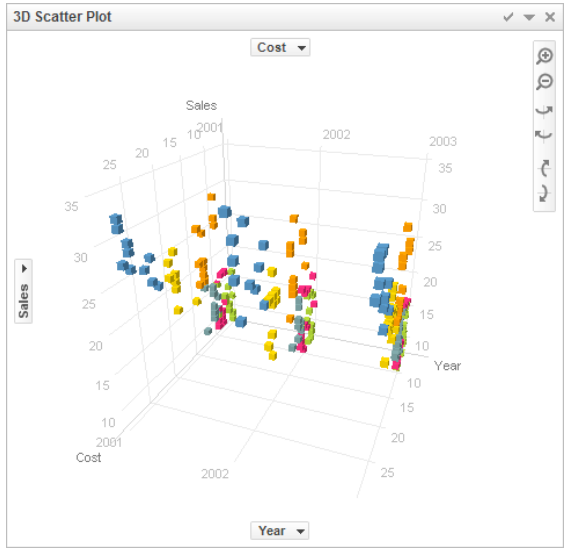


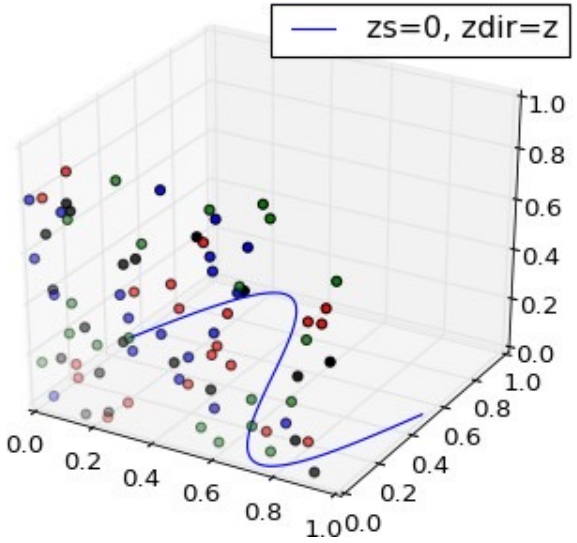
Consider using a log scale when:

- 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


X Label
Y Label
Z Label

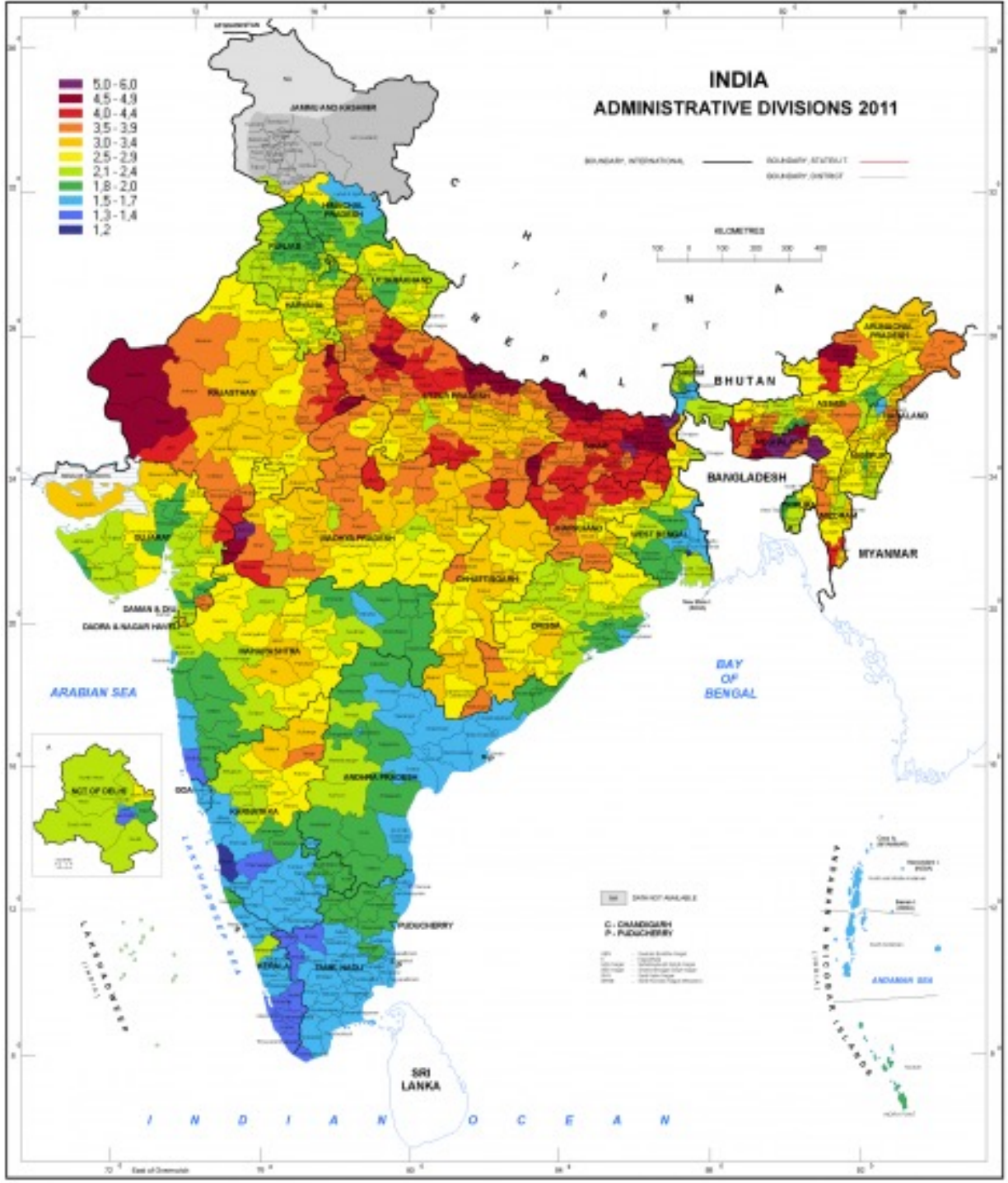
Don’t!


3D Scatter Plot
Cost
Sales
Year


zs=0, zdir=z

4. Use color strategically

Least Effective

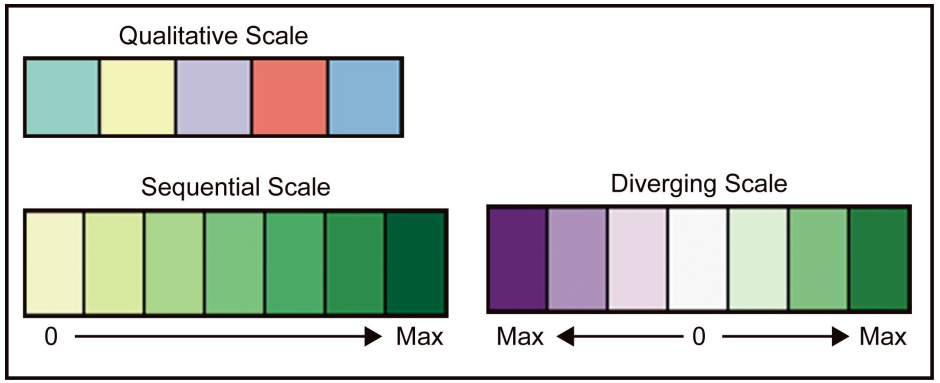


Total fertility rate map: average births per woman by districts, 2011

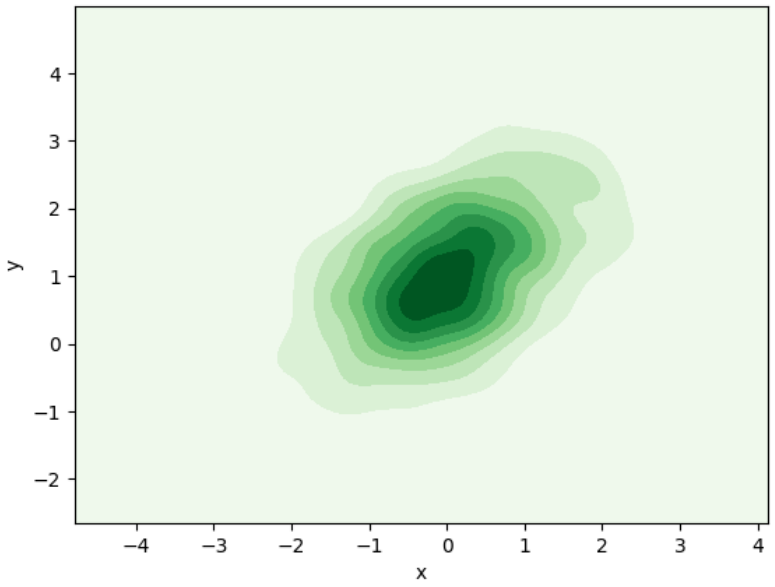
Material provided by the Data Science course at Harvard University

Nominal

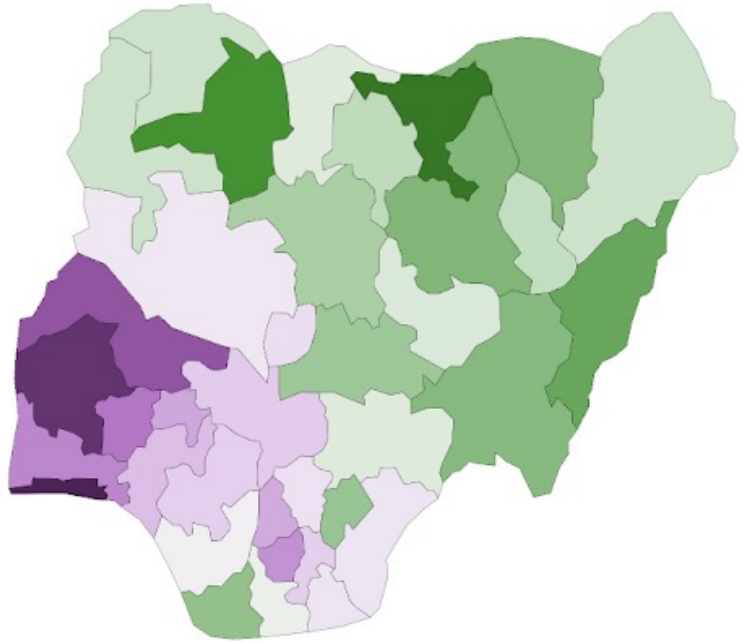
Ordinal



Ex. Densities



Ex. Correlations



Colours for Categories

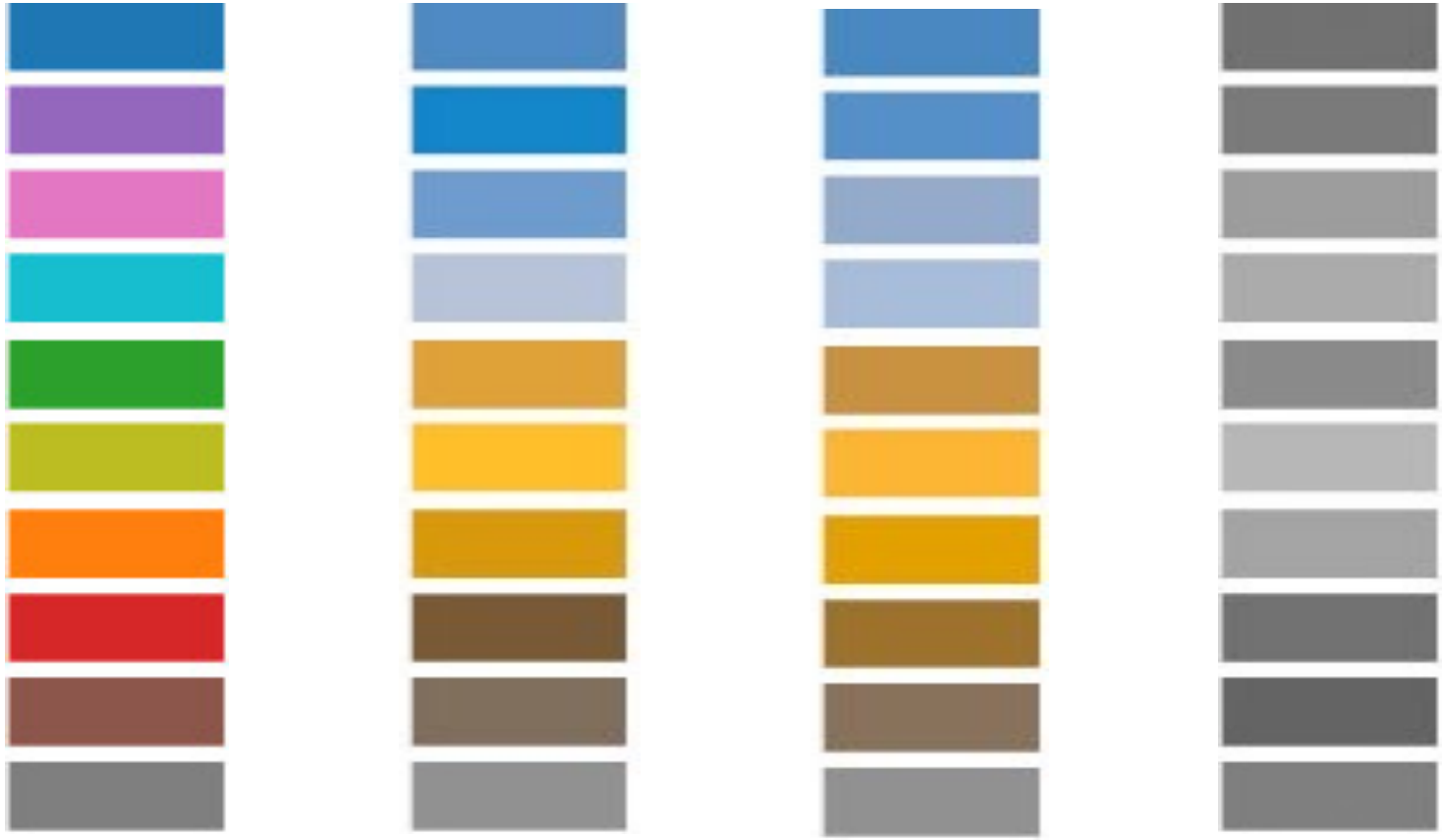
Do not use more than 5 colors at once

Colours for Ordinal Data

Vary luminance and saturation

Material provided by the Data Science course at Harvard University

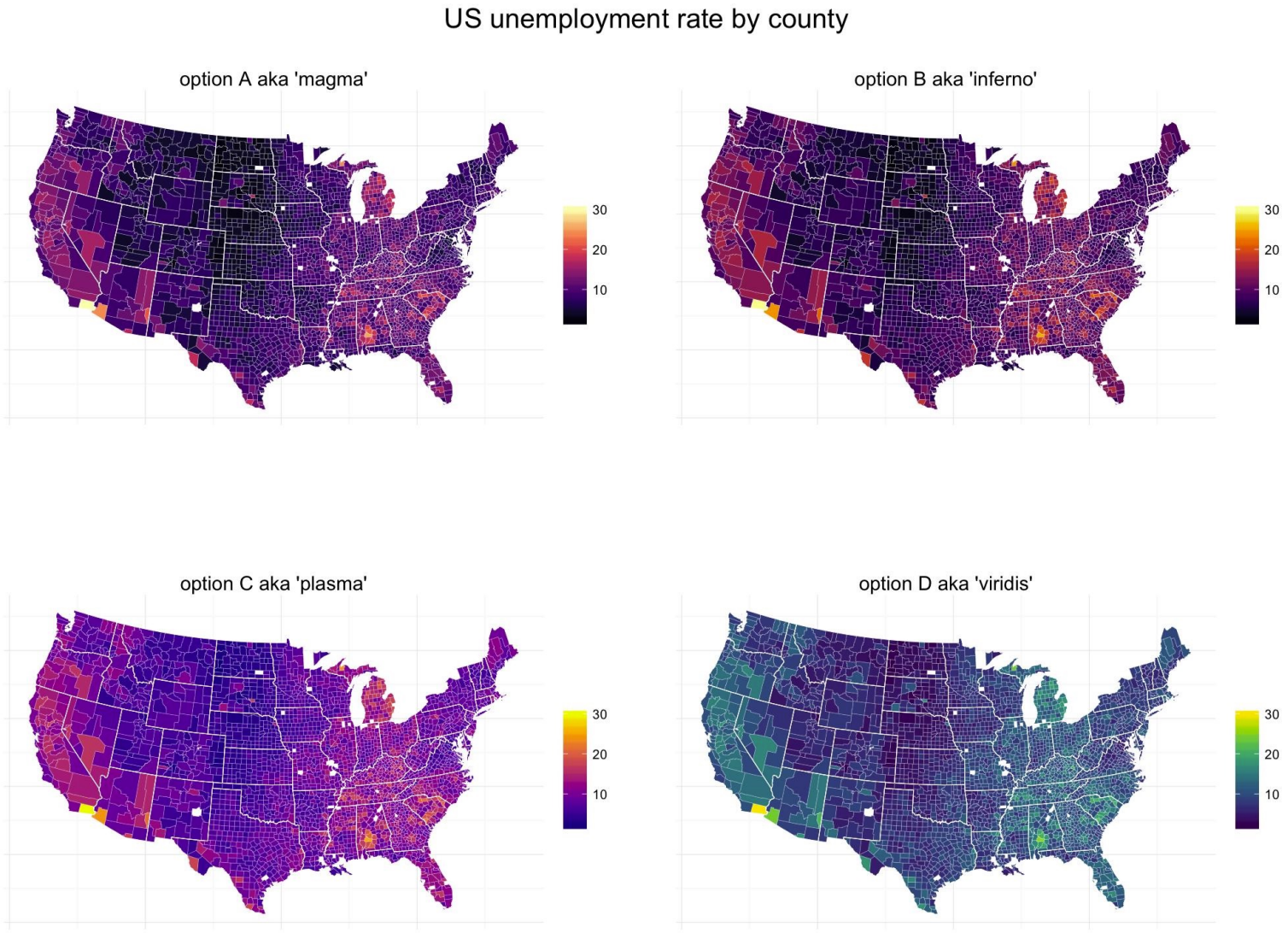
Colour Blindness



Most likely | Protanope | Deuteranope | Lightness

Material provided by the Data Science course at Harvard University

Colourmaps

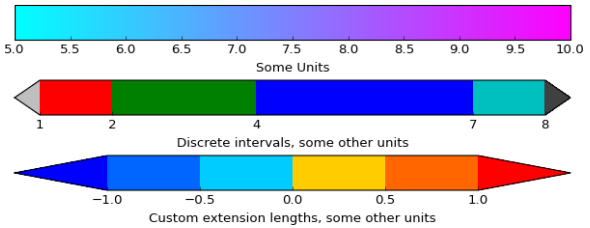


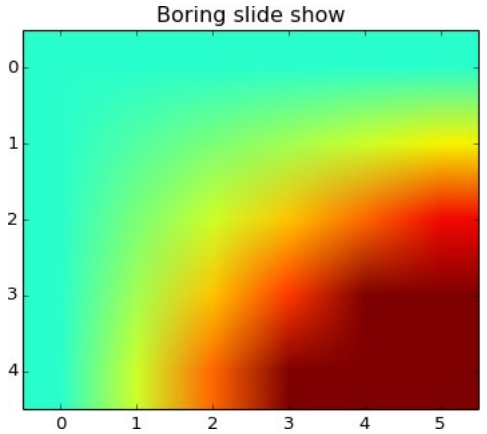
Viridis color map for better changes in perception

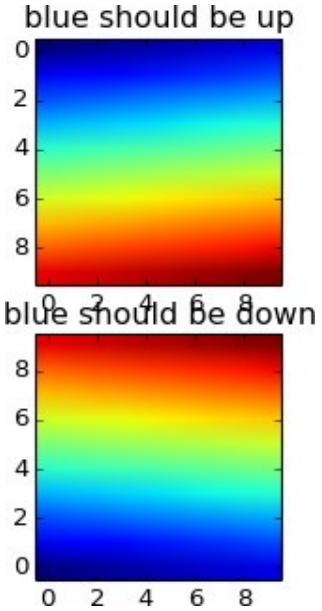
Material provided by the Data Science course at Harvard University

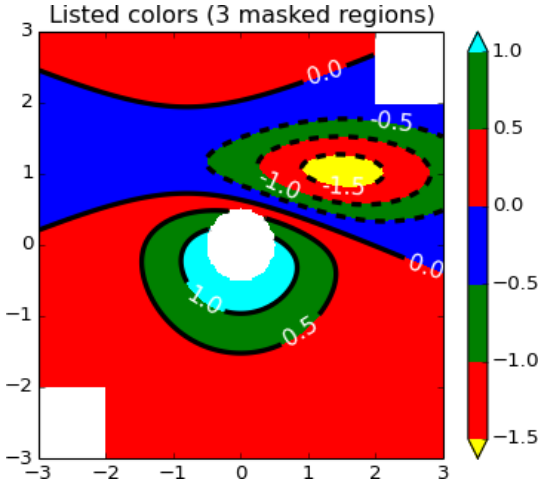
Avoid Rainbows!

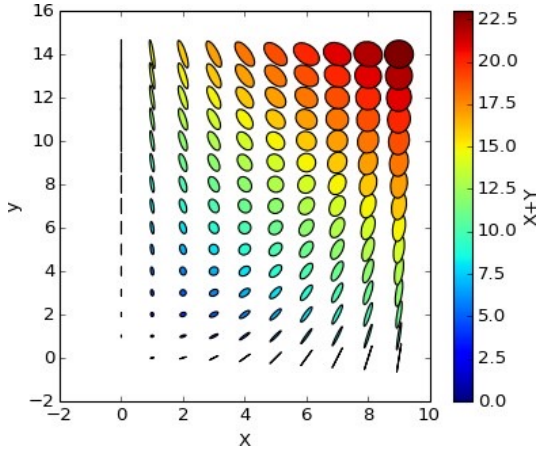
matplotlib gallery

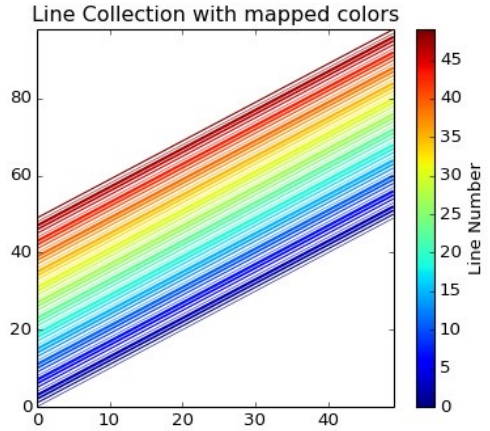












Material provided by the Data Science course at Harvard University

Checklist for Graphical Excellence

(check lecture tab for resources)

- Show the data
- Induce the viewer to think about the substance of the findings rather that 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!

Homework - Design Exercise (Time: 10 min)

In a science program, kids were asked how they feel about doing science. After the pilot program ended,

68%

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

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

Homework - 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 **Thursday 30 Nov by 2330**
5. Discussion of some of your submissions follows in Lecture 06.
6. Exercise is ***not*** graded

For next class..

Finish Labs to practice programming

Complete Homework for more practice

Check Assignment contents and due date

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