Data Visualisation

Master of Business Analytics
Melbourne Business School

ftweedie@unimelb.edu.au

Some housekeeping

Assumptions in this course

- You have the skills to munge and perform statistical analysis on your data
- You do not intend to become computer scientists or web designers
- You'll largely be working with tabular data
- You won't be working with Big Data
- You will produce visualisations for reports and presentations
- You will be producing basic visualisations for your own information
- You won't be producing visualisations for the web
- Your audience will generally be informed but not expert in the data

Scope: what's in and what's out

- Focus on static visualisations for reports and presentations
- Some interactive visualisations but the focus is not on producing interactive graphics for the web

In	**Out**
Python, R, Tableau, Kibana	Javascript, D3
Structured Data	Big Data, Unstructured Data
Graphic representations of data	Infographics

Aims

To understand when and why to visualise data

To be able to pick appropriate data visualisation styles for different types of data and different purposes

To be able to create data visualisations using a range of tools and be able to choose an appropriate tool for visualisation jobs

To use visualisation to communicate key points about a dataset

Software

Python libraries

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Plotly

Highly recommend Jupyter notebooks

R packages

Tidyverse

Installation instructions at
www.tidyverse.org/packages

Who am I?

- Qualifications in History
- Worked in policy, especially privacy and data
- Researcher training at the University of Melbourne
- Community manager at GovHack and Open Knowledge Australia
- Data scientist at The Australian Ballet
- Data and EIM at Deloitte
- Data Strategy at the University of Melbourne

Focus on communication and telling stories with data

Other stuff

Slides will be available via the LMS after each class

Accompanying Jupyter notebooks will also be available for Python and R

There will be lots of hands-on activities in classes

Datasets and a list of libraries/ packages are available via the LMS - please come to class with installation complete and the datasets somewhere you can access them

Assessment will be a small-group presentation. Details and data for the assignment are available on the LMS

Visualising data

What is data visualisation

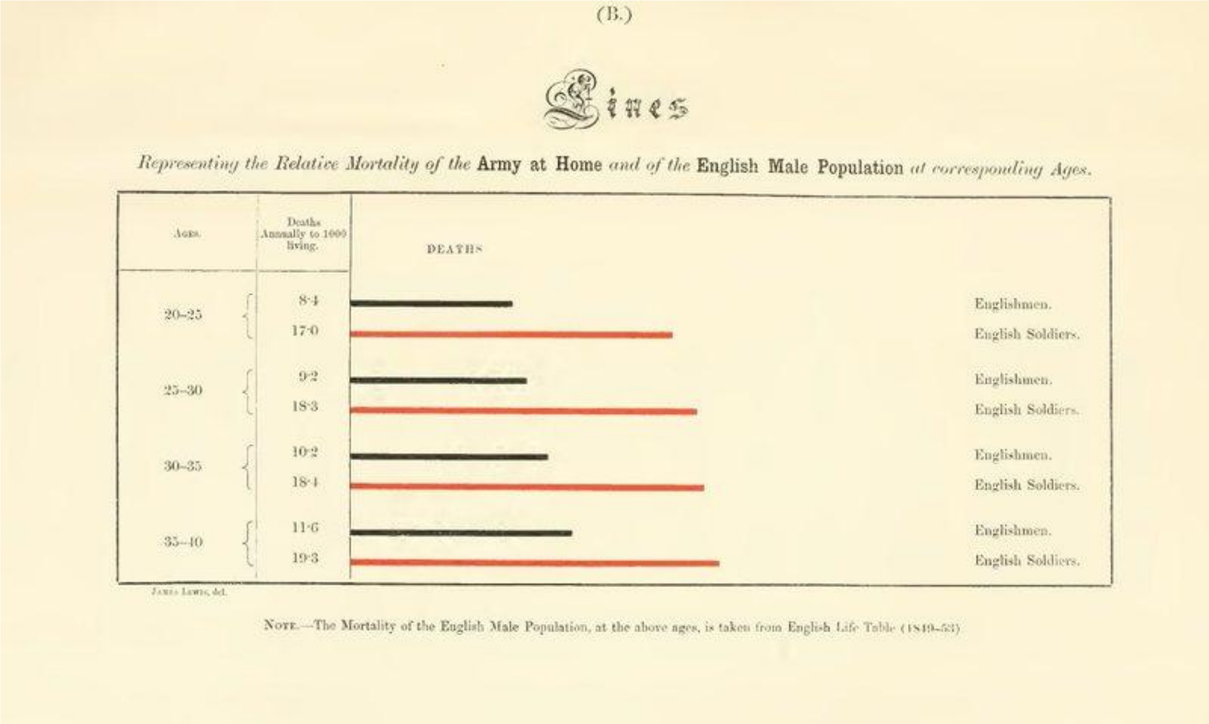
Data visualisation involves the creation and study of **visual representations of data** to communicate information clearly and effectively

Data visualisation should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production or something else
- avoid distorting what the data has to say
- present many numbers in a small space
- make large data sets 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 reasonably clear purpose: description, exploration, tabulation or decoration
- be closely integrated with the statistical and verbal descriptions of a data set.

Edward Tufte *The Visual Display of Quantitative Information*

Florence Nightingale: Saving lives with data



(B.)

Lines

Representing the Relative Mortality of the **Army at Home** *and of the* **English Male Population** *at corresponding Ages.*

Note.—The Mortality of the English Male Population, at the above ages, is taken from English Life Table (1849–53)

assets.atlasobscura.com/article_images/42024/image.jpg

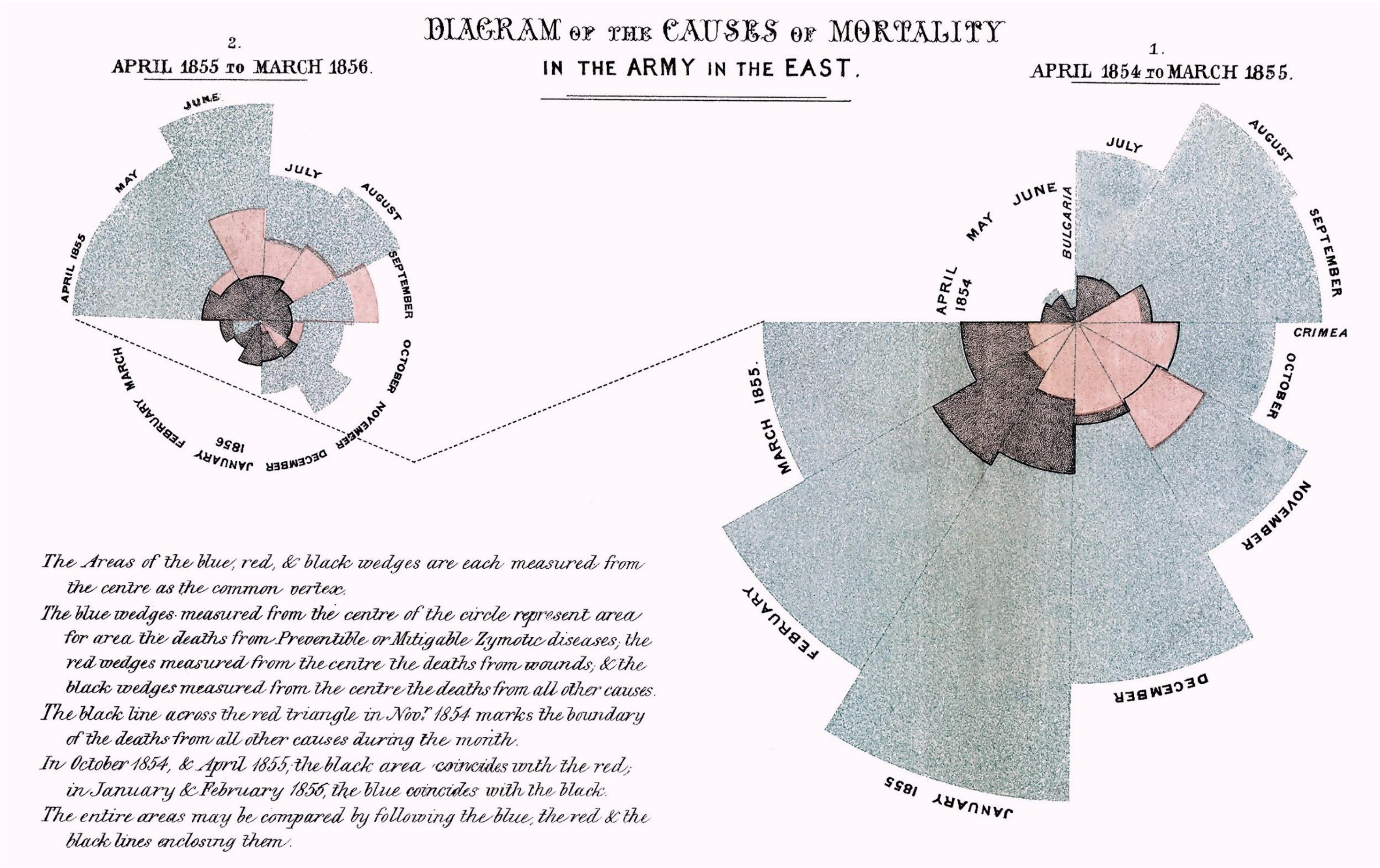


DIAGRAM OF THE CAUSES OF MORTALITY IN THE ARMY IN THE EAST.

The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Novr. 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

upload.wikimedia.org/wikipedia/commons/1/17/Nightingale-mortality.jpg

Why visualise data

Explore

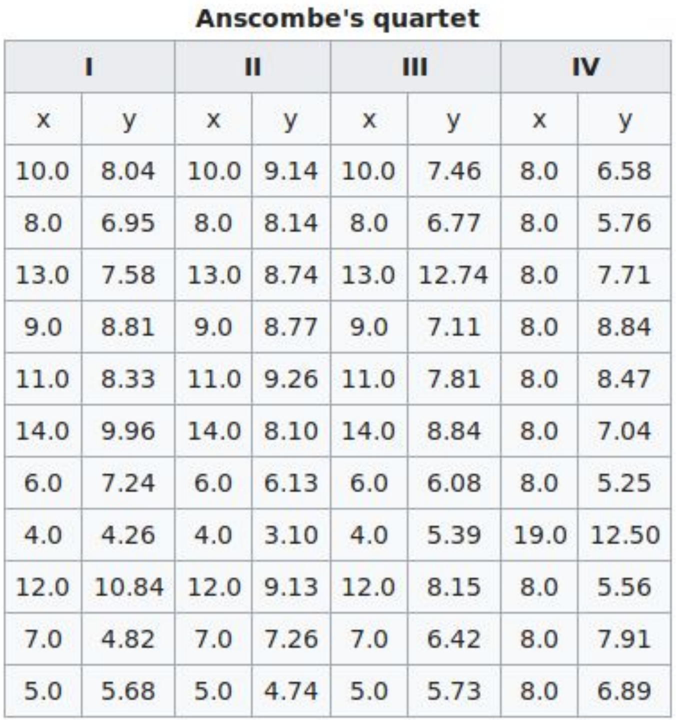
Discuss (Educate)

Decide (Persuade)

Audience: You

Someone else

More than summary statistics

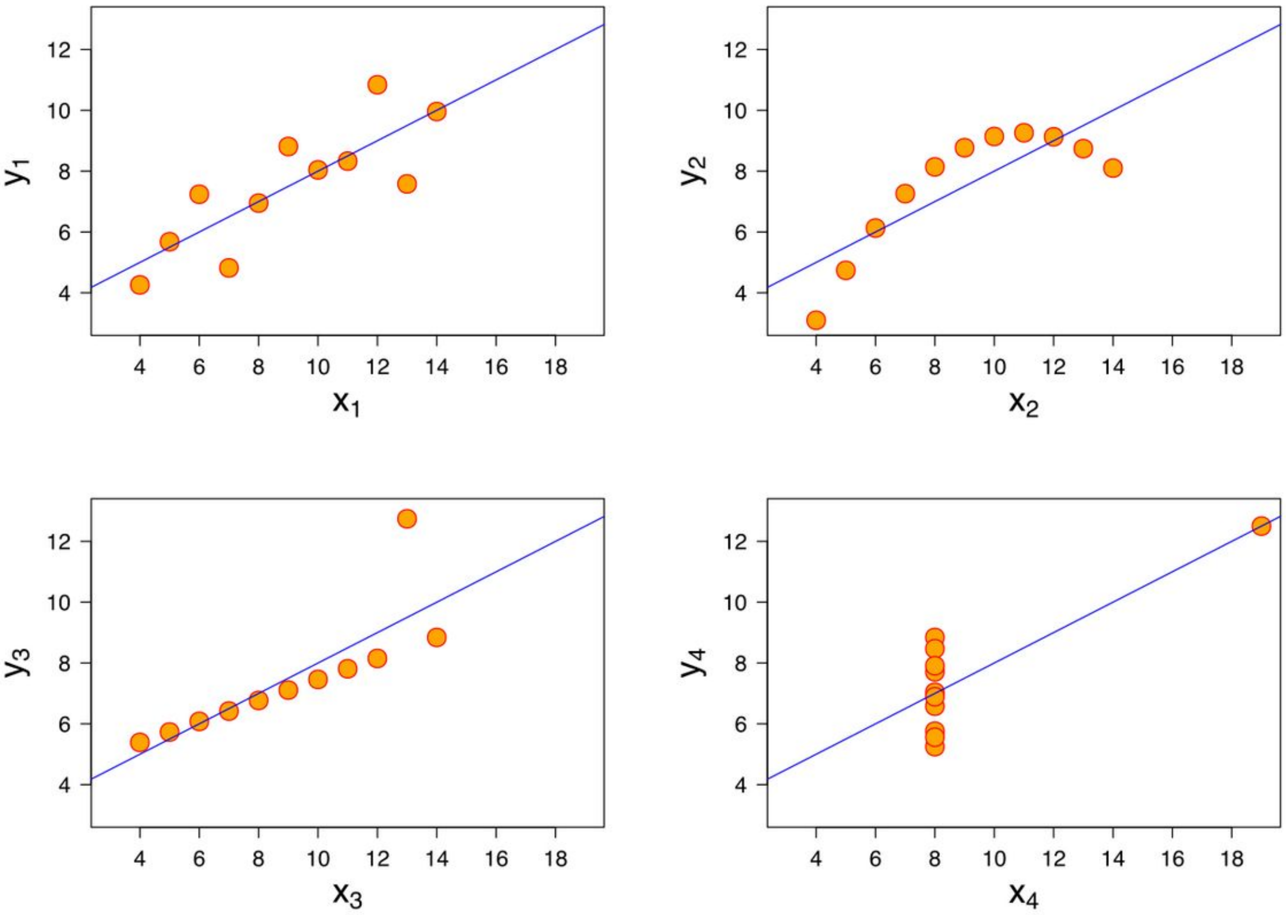


Anscombe's quartet

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

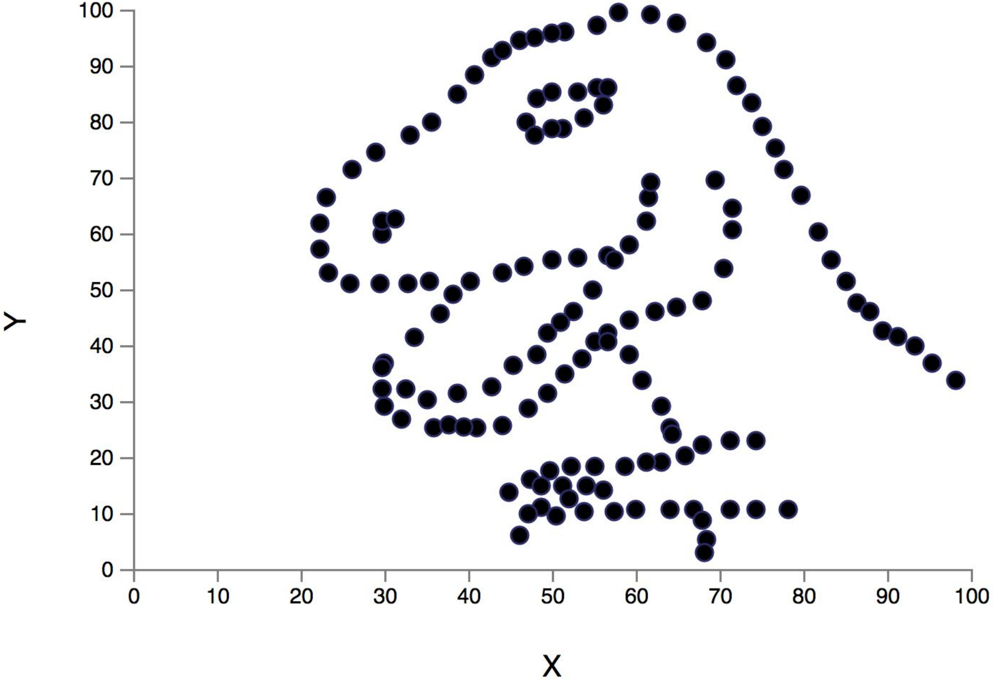
en.wikipedia.org/wiki/Anscombe%27s_quartet

Property	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.5
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	y = 3.00 + 0.5x
Coefficient of determination of linear regression	0.67



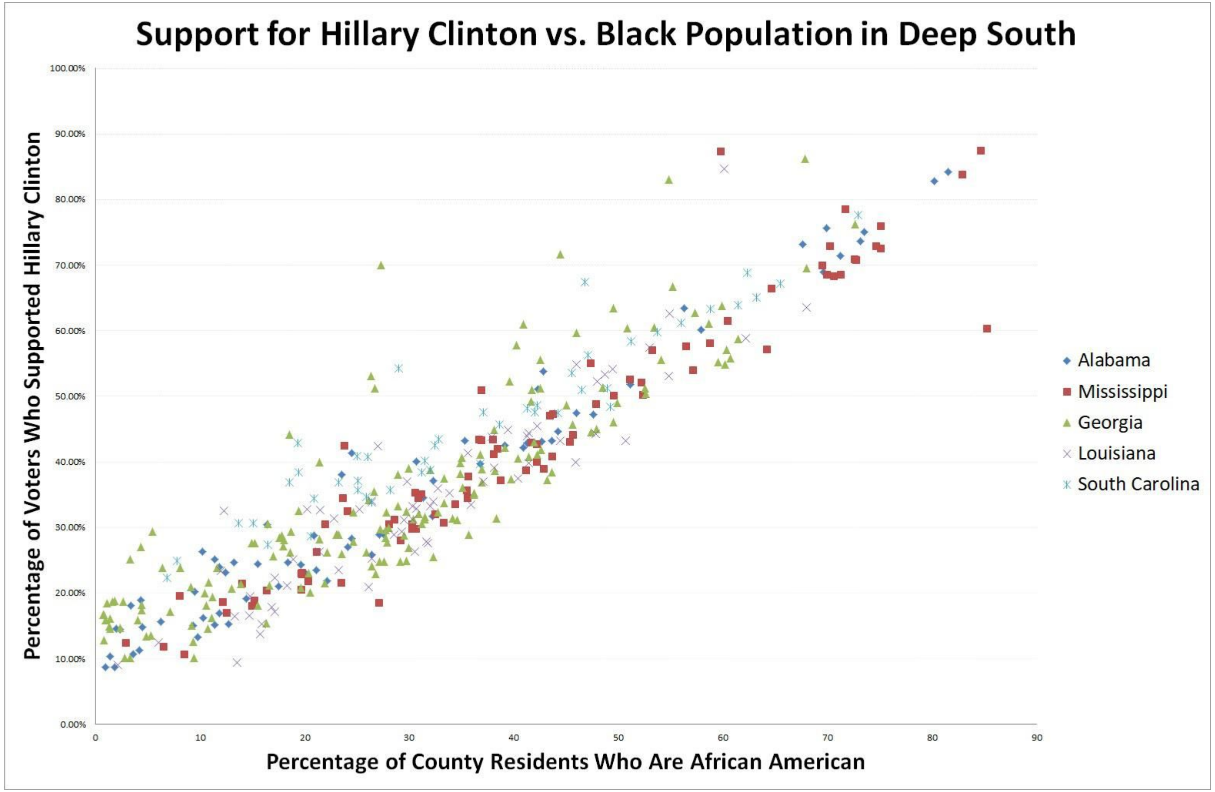
en.wikipedia.org/wiki/Anscombe%27s_quartet

N	142
X mean	54.2633
X SD	16.7651
Y mean	47.8323
Y SD	26.9354
Pearson Correlation	-0.0645



thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html

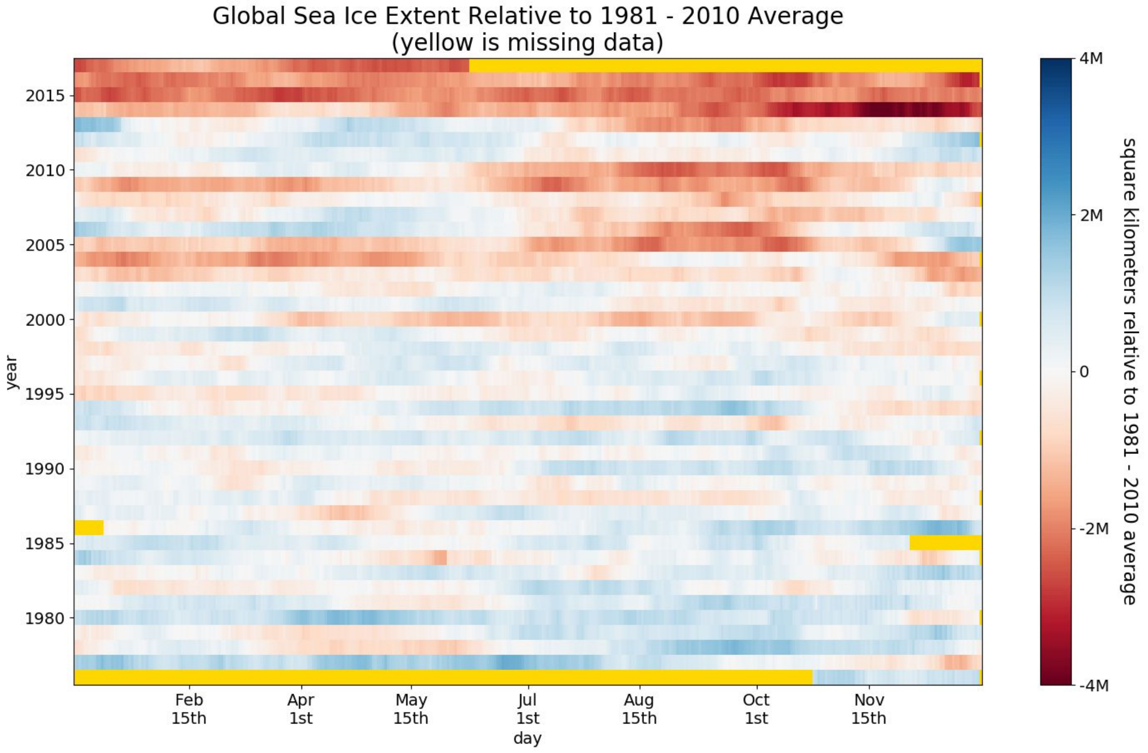
Discuss



reddit.com/r/dataisbeautiful/comments/c0cfuv/countybycounty_results_from_the_2016_us/

Decide

Global sea ice change over the past 40 years



reddit.com/r/dataisbeautiful/comments/byxxbm/global_sea_ice_change_over_the_past_40_years_oc/

Data visualisation should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production or something else
- avoid distorting what the data has to say
- present many numbers in a small space
- make large data sets 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 reasonably clear purpose: description, exploration, tabulation or decoration
- be closely integrated with the statistical and verbal descriptions of a data set.

Edward Tufte *The Visual Display of Quantitative Information*

Looking at Data

Types of data

There are various ways of talking about types of data

Quantitative	
Discrete	Continuous

Qualitative
Categorical

Generally, you can’t perform mathematical transformations on categorical data

Categorical/ Nominal

Items that are differentiated by name or category. Categories are distinct. They may be grouped but can’t be mathematically altered

For example:

- Gender
- Country of birth
- Type of pet
- Colour

Ordinal

Ordinal data describes data points relative to each other. The sequence is important, but the distance between categories is not necessarily fixed or known

For example:

- Finishing order in a race
- Salary bands
- Likert scale (strongly agree, agree, neutral, disagree, strongly disagree)

Interval/ Integer

Measured along a continuous scale in which each position is equidistant from one another. This allows for the distance between two pairs to be equivalent in some way. Generally can’t be multiplied or divided

For example:

- Degrees celsius
- Date

Ratio

Numbers can be compared as multiples of one another and zero has meaning. The interval between measures is consistent. Specifies “how much” or “how many”

For example:

- Mass
- Length
- Duration
- Cost

Discrete v Continuous

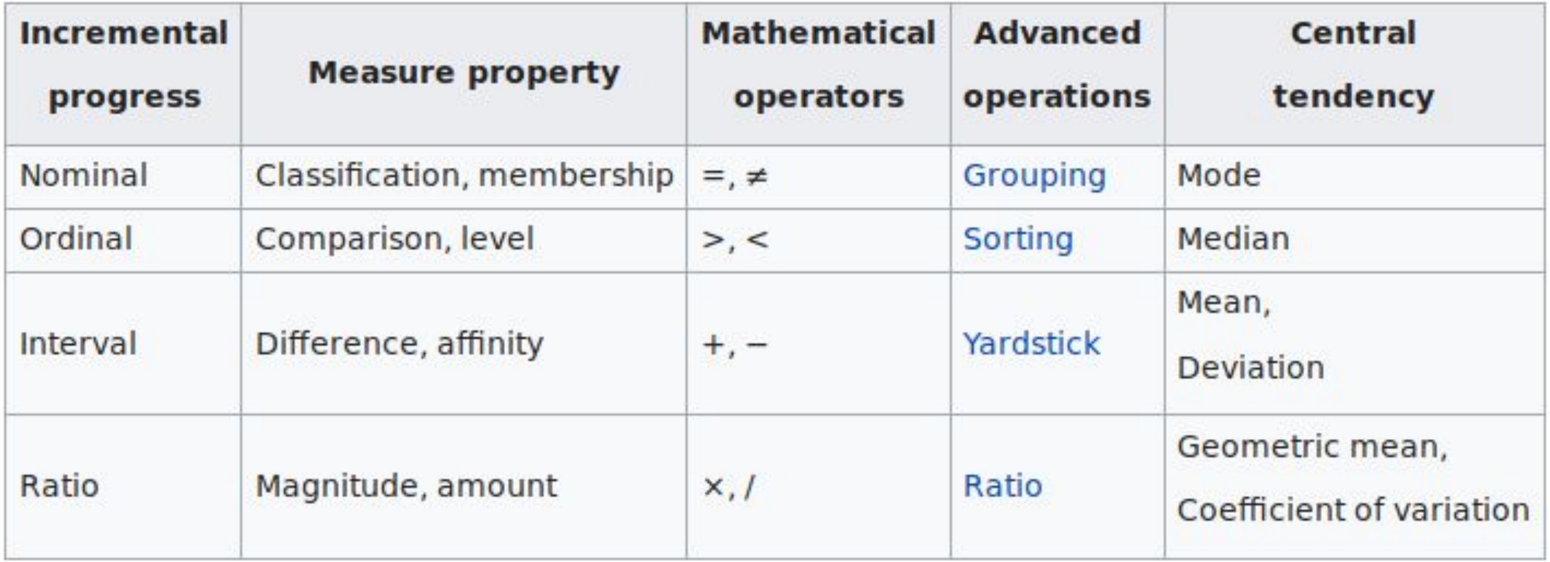
Continuous measures are measured along a continuous scale which can be divided into fractions, such as temperature. Continuous variables allow for infinitely fine sub-division, which means if you can measure sufficiently accurately, you can compare two items and determine the difference.

Discrete variables are measured across a set of fixed values.

Exercise: Donate some data

https://forms.gle/zqbEFgFnEPjh4gwF8

Stevens' Typology



Incremental progress	Measure property	Mathematical operators	Advanced operations	Central tendency
Nominal	Classification, membership	=, ≠	Grouping	Mode
Ordinal	Comparison, level	>, <	Sorting	Median
Interval	Difference, affinity	+, −	Yardstick	Mean, Deviation
Ratio	Magnitude, amount	×, /	Ratio	Geometric mean, Coefficient of variation

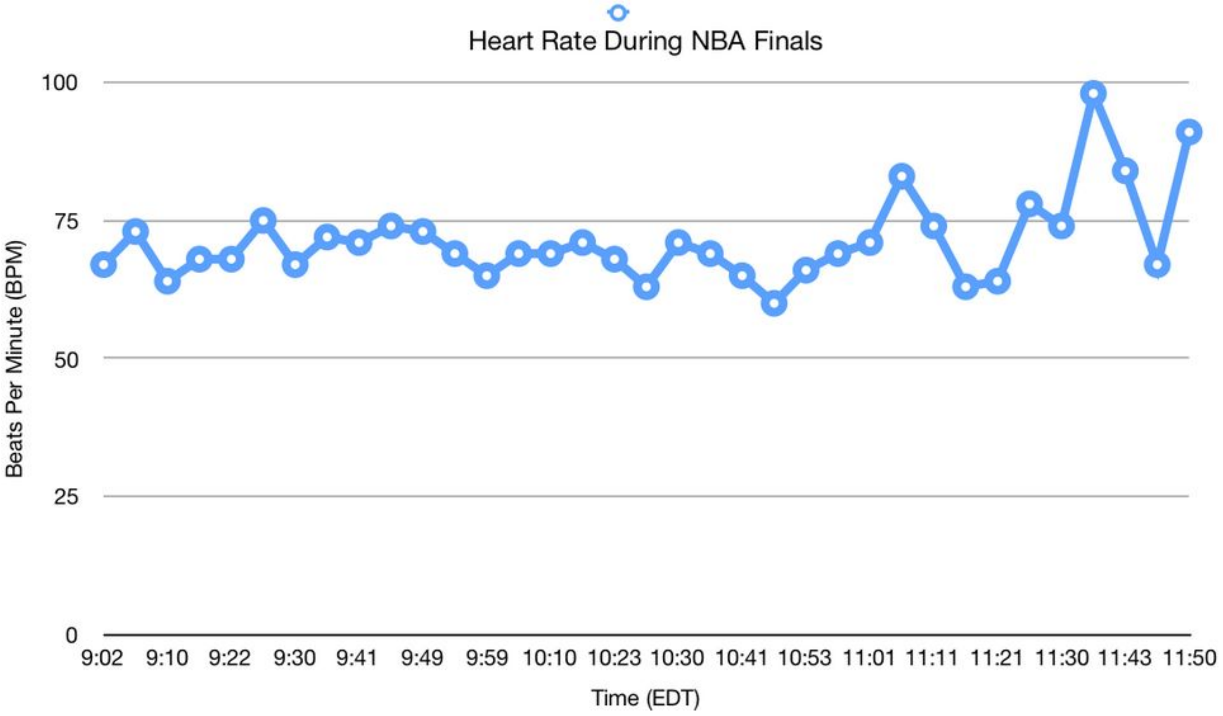
en.wikipedia.org/wiki/Level_of_measurement#Interval_scale

Data Vis can be used to show

1. Change over time
2. Ranking
3. Proportion (part to whole)
4. Deviation
5. Frequency distribution
6. Correlation
7. Categorical comparison
8. Geographic or geospatial

Change over time

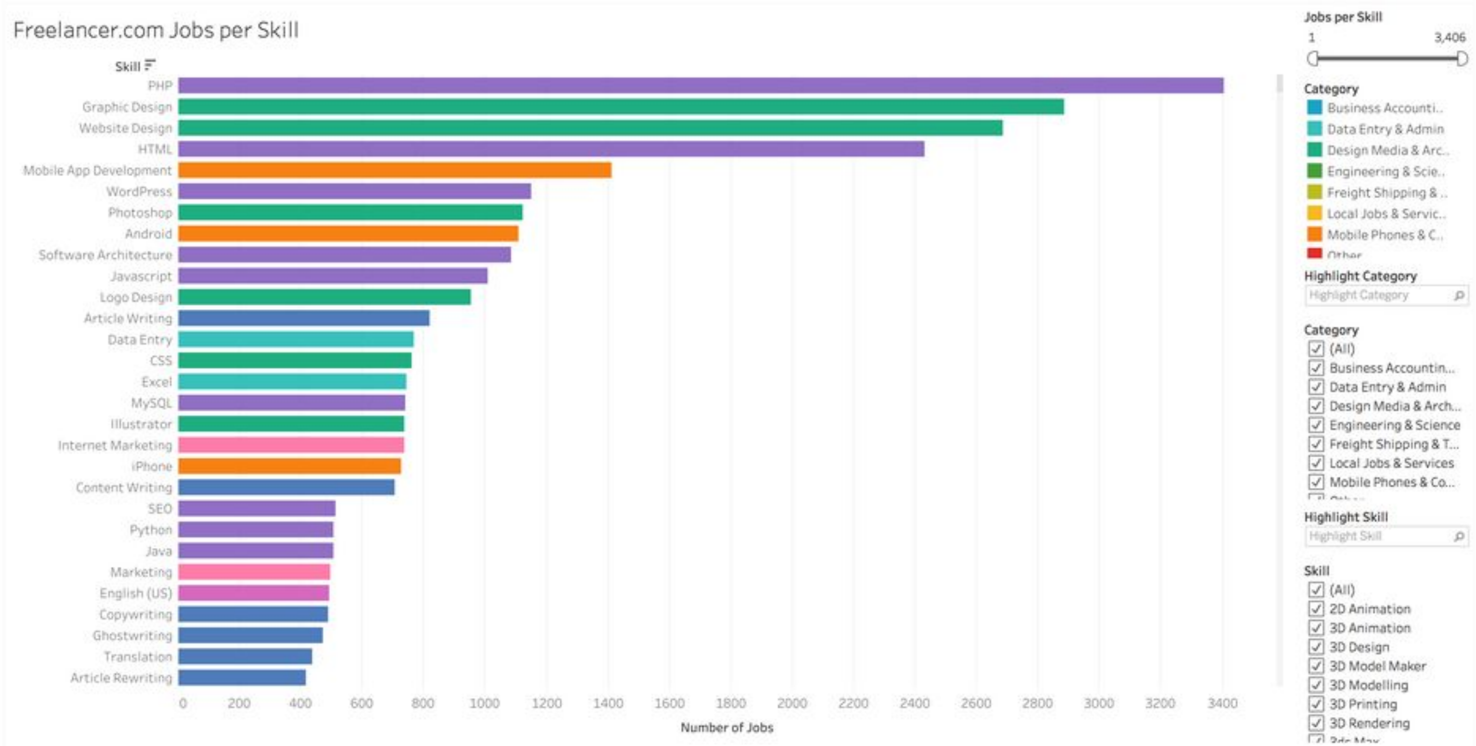
Variables are tracked over a period of time



reddit.com/r/dataisbeautiful/comments/c0qzra/oc_my_heart_rate_while_watching_the_nba_finals/

Ranking

Categories are ranked



Jobs per Skill

1 3,406

Category

☑ (All)
☑ Business Accountin...
☑ Data Entry & Admin
☑ Design Media & Arch...
☑ Engineering & Science
☑ Freight Shipping & T...
☑ Local Jobs & Services
☑ Mobile Phones & Co...

Skill

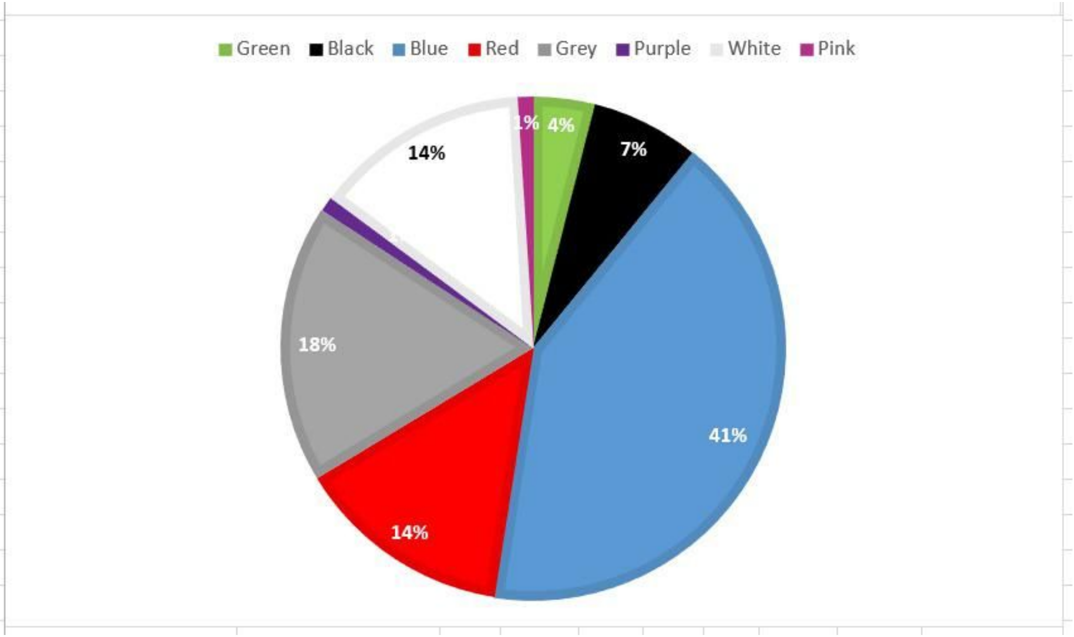
☑ (All)
☑ 2D Animation
☑ 3D Animation
☑ 3D Design
☑ 3D Model Maker
☑ 3D Modelling
☑ 3D Printing
☑ 3D Rendering

reddit.com/r/dataisbeautiful/comments/c017c9/the_number_of_job_postings_for_the_top_1000/

Proportion

Categorical subdivisions presented as a proportion of the whole

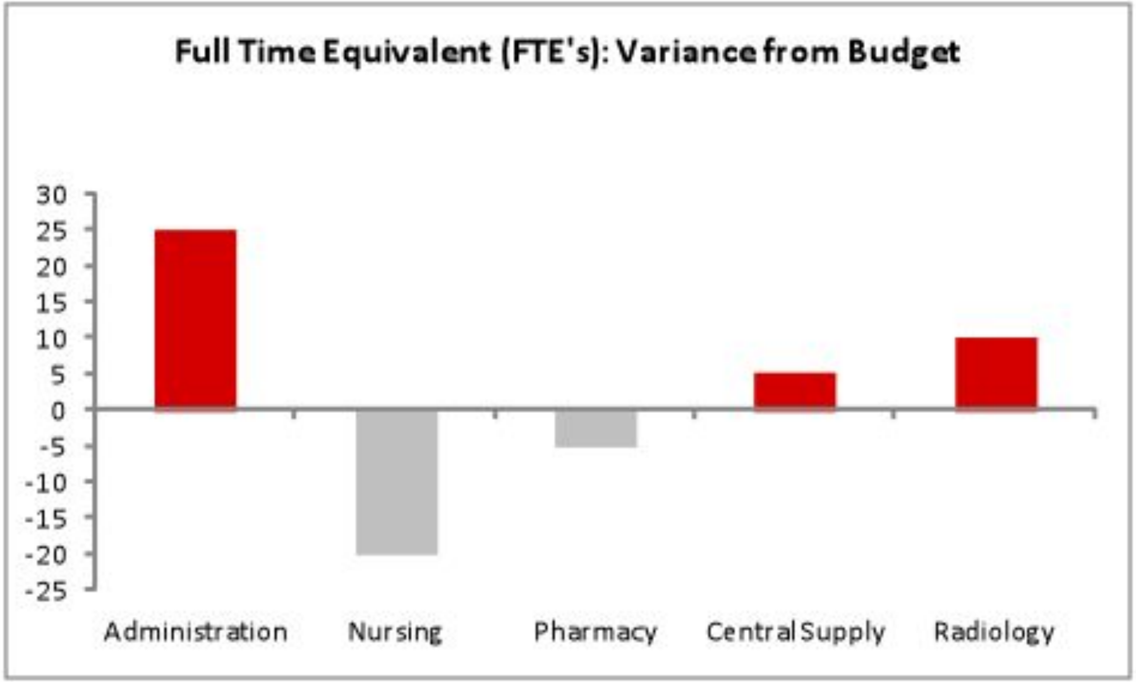
My Boss’s Shirt Colour



reddit.com/r/dataisbeautiful/comments/byq9fs/oc_my_bosss_shirt_color/

Deviation

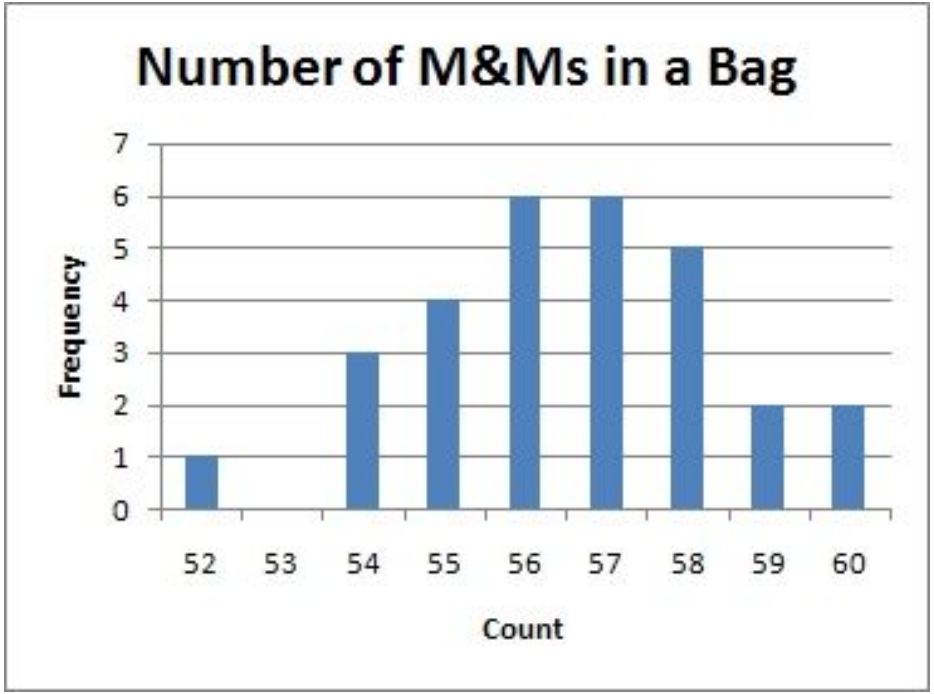
Categories are compared against a reference (e.g. actual vs budget)



ksrowell.com/blog-visualizing-data/2013/02/14/how-to-make-a-deviation-graph/

Frequency Distribution

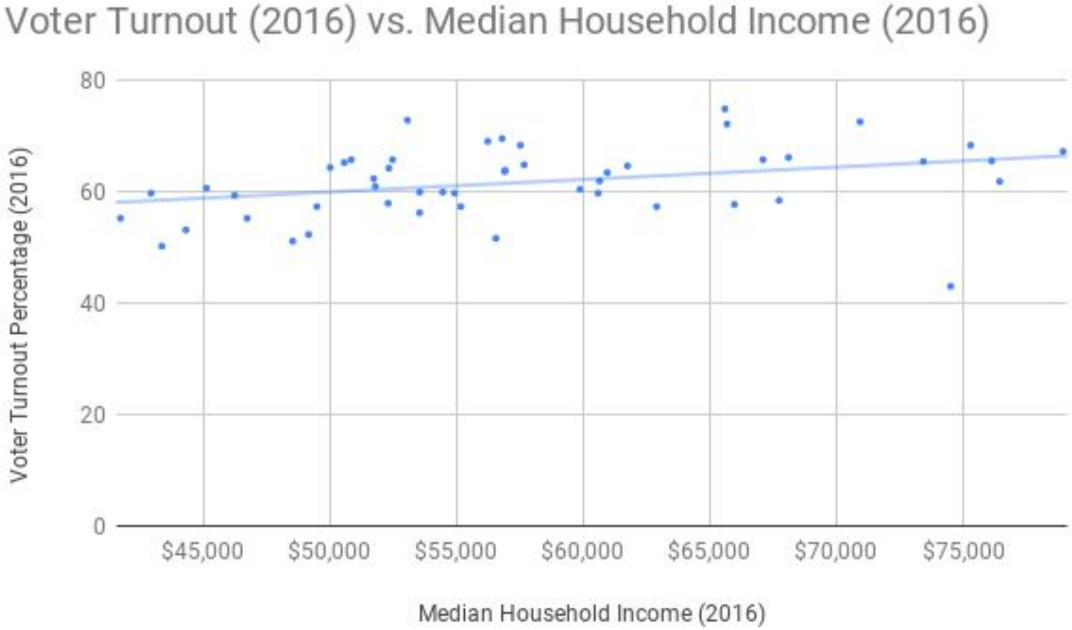
Number of observations of a particular variable for a given interval



dsearls.org/courses/M120Concepts/ClassNotes/Statistics/lab1c.jpg

Correlation

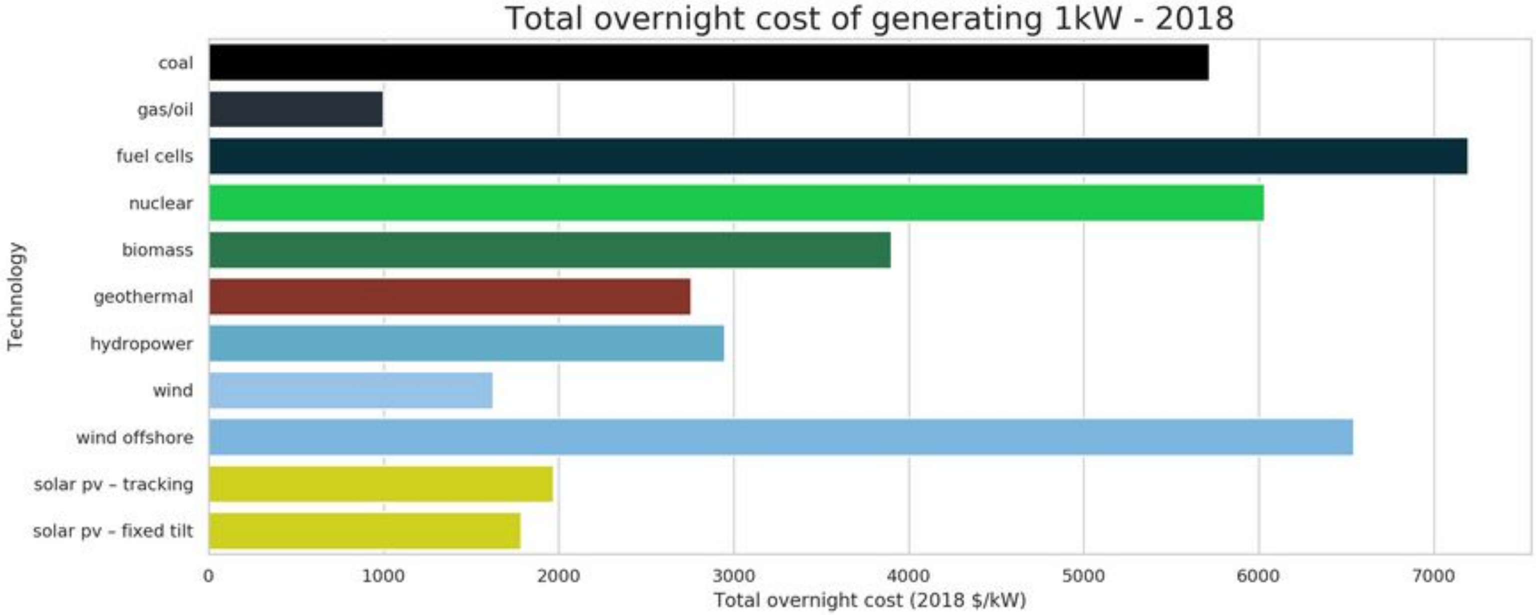
Comparison between two variables to determine if they are related



.reddit.com/r/dataisbeautiful/comments/c0eg64/voter_turnout_in_the_2016_us_presidential/

Categorical comparison

Compares categories in no particular order (distinct from ranking, which does have an order)

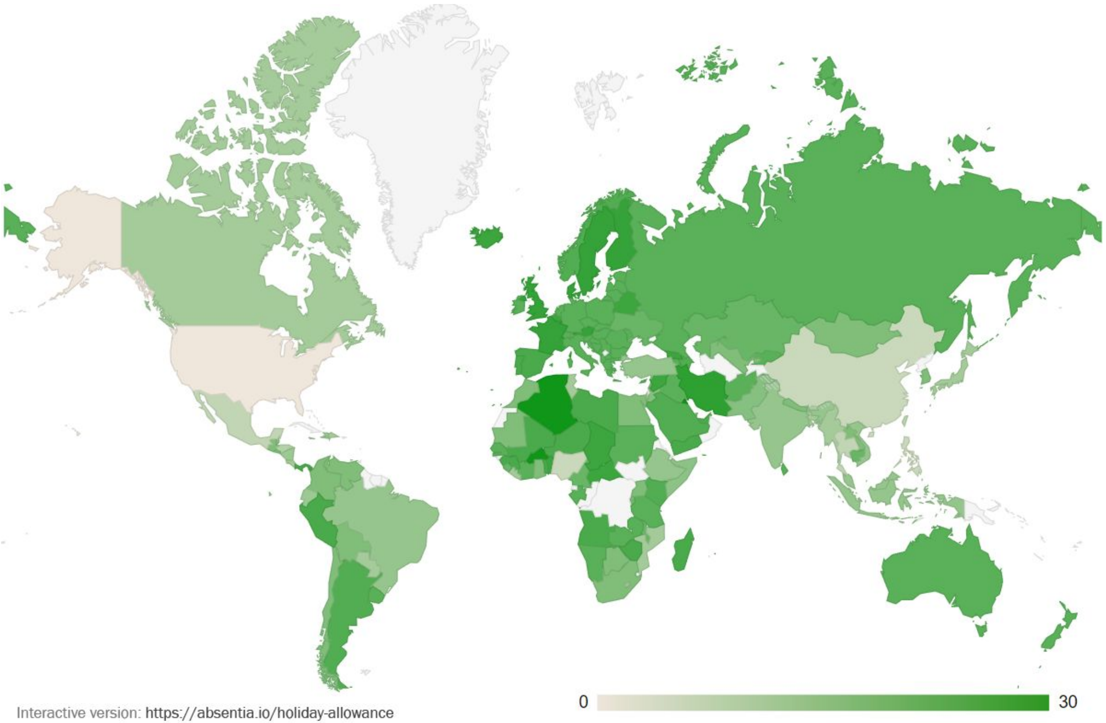


reddit.com/r/dataisbeautiful/comments/bvk1fq/oc_cost_of_generating_1kw_from_different_sources/

Geospatial

Comparison of a variable by a spatial category

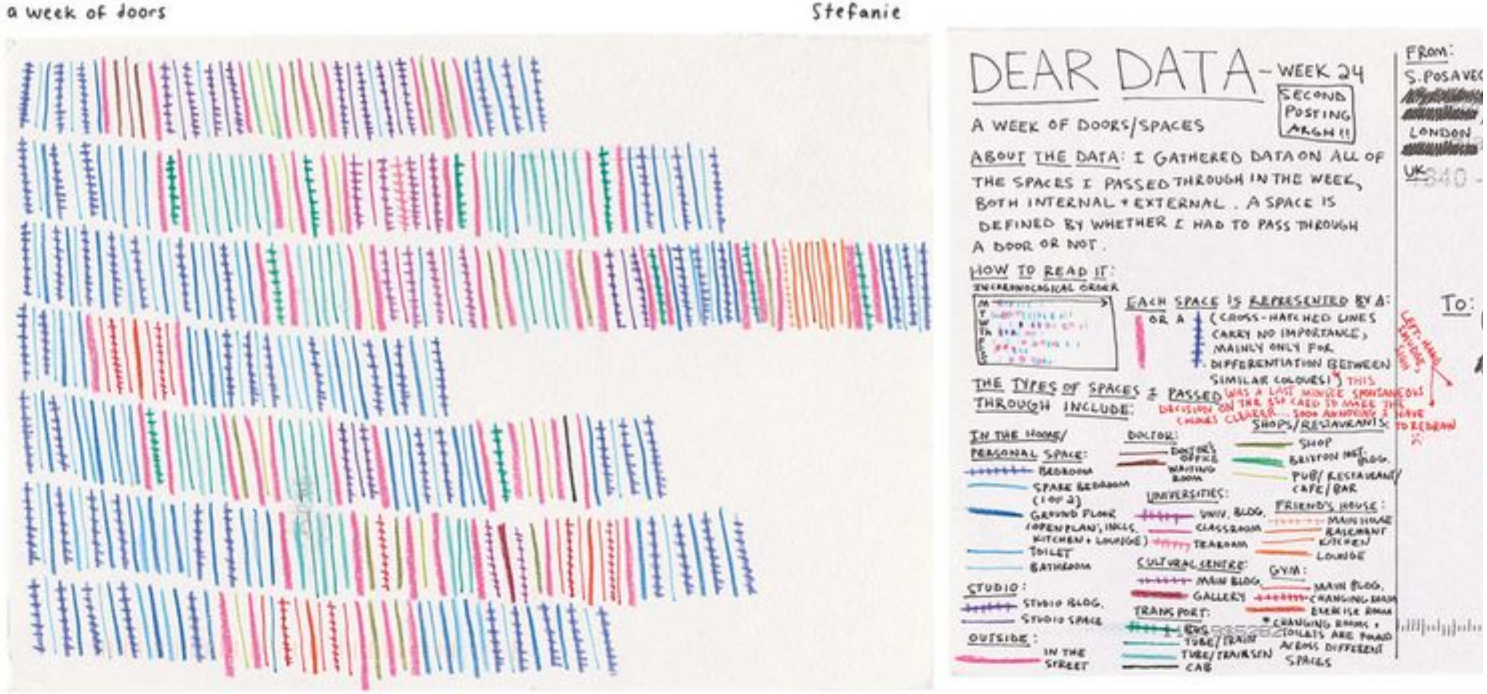
Minimum annual leave by country



Interactive version: https://absentia.io/holiday-allowance

Dear Data

https://www.dear-data.com/theproject



DEAR DATA - WEEK 24

SECOND POSTING ARGH!!

A WEEK OF DOORS/SPACES

ABOUT THE DATA: I GATHERED DATA ON ALL OF THE SPACES I PASSED THROUGH IN THE WEEK, BOTH INTERNAL + EXTERNAL. A SPACE IS DEFINED BY WHETHER I HAD TO PASS THROUGH A DOOR OR NOT.

HOW TO READ IT:
IN CHRONOLOGICAL ORDER

M
T
W
Th
F
S
S

EACH SPACE IS REPRESENTED BY A: | OR A ‡ (CROSS-HATCHED LINES CARRY NO IMPORTANCE, MAINLY ONLY FOR DIFFERENTIATION BETWEEN SIMILAR COLOURS!) THIS WAS A LAST MINUTE SPONTANEOUS DECISION ON THE 1ST CARD TO MAKE THE CHOICES CLEARER... SOOO ANNOYING I HAVE TO REDRAW :(

LEFT-HAND SMUDGE, UGH

THE TYPES OF SPACES I PASSED THROUGH INCLUDE:

IN THE HOME/ PERSONAL SPACE:
BEDROOM
SPARE BEDROOM (1 OF 2)
GROUND FLOOR (OPEN PLAN, INCLS. KITCHEN + LOUNGE)
TOILET
BATHROOM

STUDIO:
STUDIO BLDG.
STUDIO SPACE

OUTSIDE:
IN THE STREET

DOCTOR:
DOCTOR'S OFFICE
WAITING ROOM

UNIVERSITIES:
UNIV. BLDG.
CLASSROOM
TEAROOM

CULTURAL CENTRE:
MAIN BLDG.
GALLERY

TRANSPORT:
BUS
TUBE/TRAIN
TUBE/TRAIN STN
CAB

SHOPS/RESTAURANTS:
SHOP
BRIXTON MKT. BLDG.
PUB/ RESTAURANT/ CAFE/ BAR

FRIEND'S HOUSE:
MAIN HOUSE
BASEMENT
KITCHEN
LOUNGE

GYM:
MAIN BLDG.
CHANGING ROOM*
EXERCISE ROOM
* CHANGING ROOMS + TOILETS ARE FOUND ACROSS DIFFERENT SPACES

FROM:
S. POSAVEC
[redacted]
[redacted]
LONDON
[redacted]
UK

TO:

Exercise: Your day as data

Create five graphics that tell a story about your day or week

1. Change over time
2. Ranking
3. Proportion (part to whole)
4. Deviation
5. Frequency distribution
6. Correlation
7. Categorical comparison
8. Geographic or geospatial

Representing Data

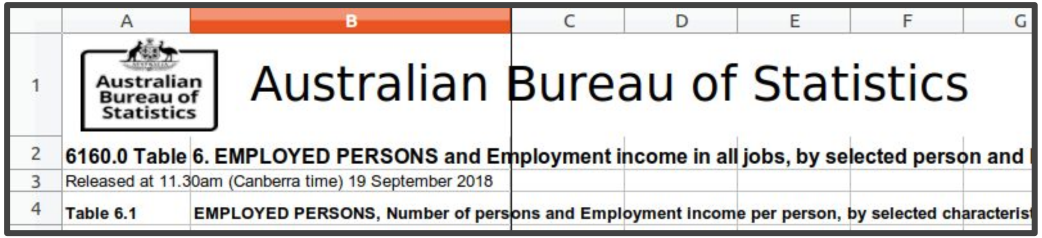
Side note: Tidy data

To be able to programatically analyse data, it needs to be tidy!

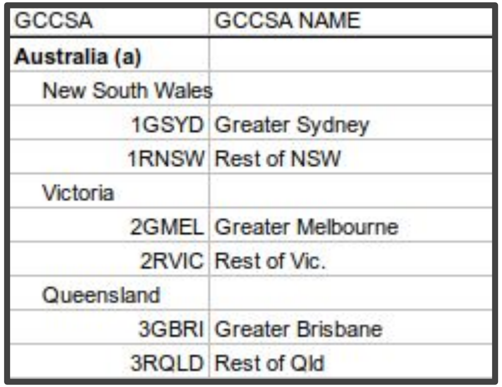
Tidy data has one variable per column and one observation per row

A tidy spreadsheet has column names in the top row

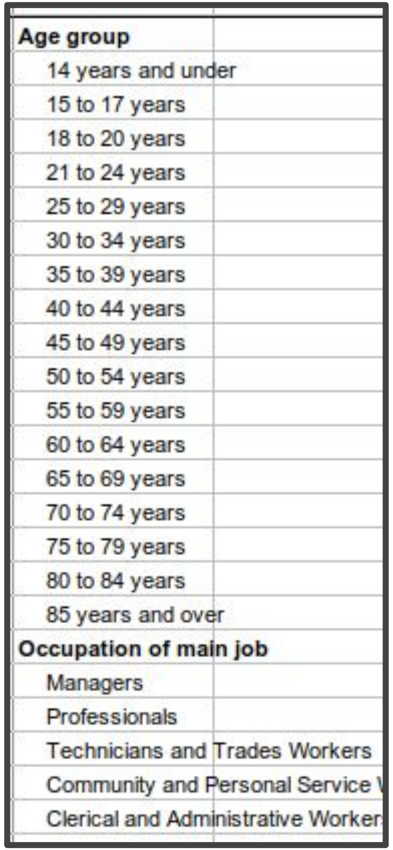
Tidy data has one data type and unit of measurement in each column



	A	B	C	D	E	F	G
1	Australian Bureau of Statistics	Australian Bureau of Statistics					
2	6160.0 Table 6. EMPLOYED PERSONS and Employment income in all jobs, by selected person and						
3	Released at 11.30am (Canberra time) 19 September 2018						
4	Table 6.1	EMPLOYED PERSONS, Number of persons and Employment income per person, by selected characterist					

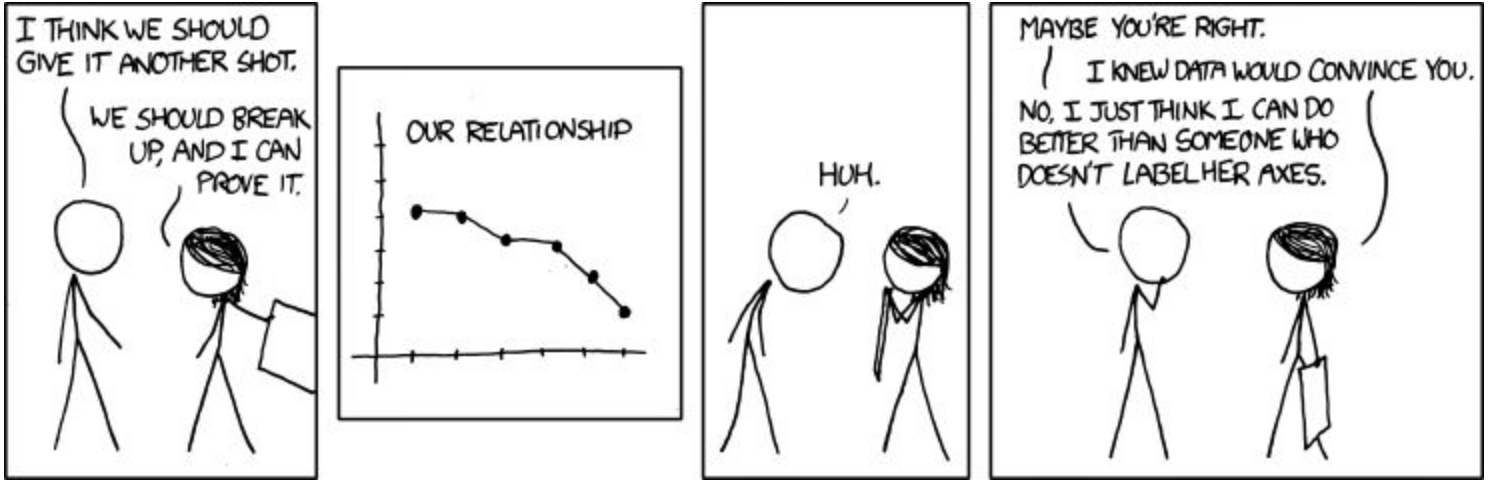


GCCSA	GCCSA NAME
Australia (a)	
New South Wales	
1GSYD	Greater Sydney
1RNSW	Rest of NSW
Victoria	
2GMEL	Greater Melbourne
2RVIC	Rest of Vic.
Queensland	
3GBRI	Greater Brisbane
3RQLD	Rest of Qld



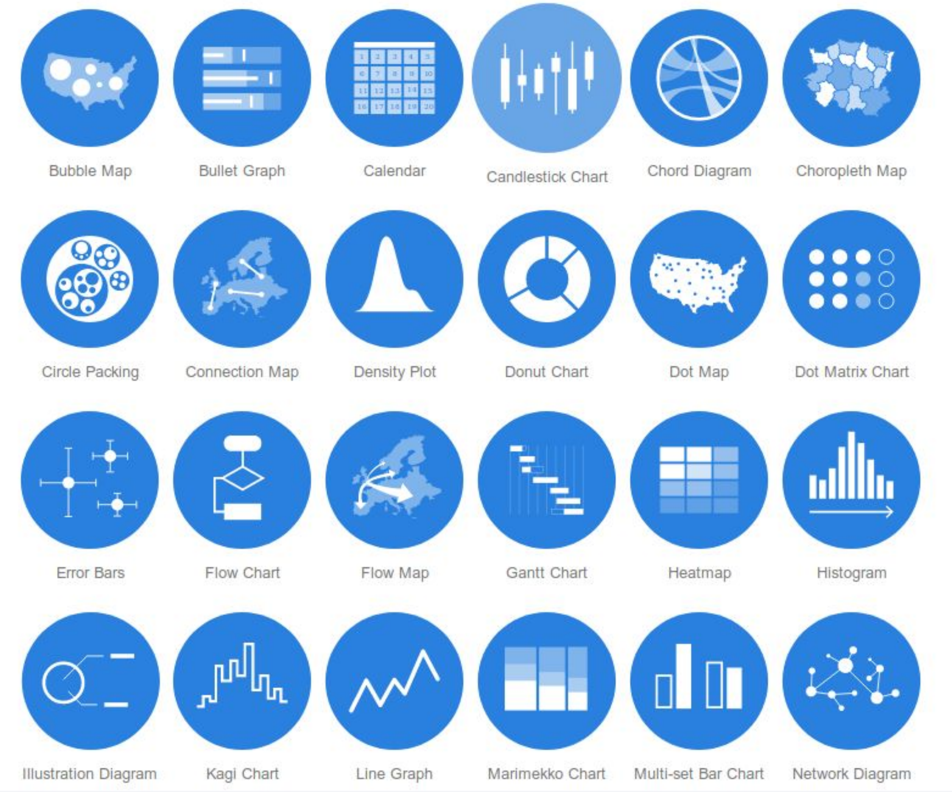
Age group	
14 years and under	
15 to 17 years	
18 to 20 years	
21 to 24 years	
25 to 29 years	
30 to 34 years	
35 to 39 years	
40 to 44 years	
45 to 49 years	
50 to 54 years	
55 to 59 years	
60 to 64 years	
65 to 69 years	
70 to 74 years	
75 to 79 years	
80 to 84 years	
85 years and over	
Occupation of main job	
Managers	
Professionals	
Technicians and Trades Workers	
Community and Personal Service	
Clerical and Administrative Worker	

Basic good practice


I THINK WE SHOULD GIVE IT ANOTHER SHOT.
WE SHOULD BREAK UP, AND I CAN PROVE IT.
OUR RELATIONSHIP
HUH.
MAYBE YOU'RE RIGHT.
I KNEW DATA WOULD CONVINCE YOU.
NO, I JUST THINK I CAN DO BETTER THAN SOMEONE WHO DOESN'T LABEL HER AXES.

https://www.xkcd.com/833/

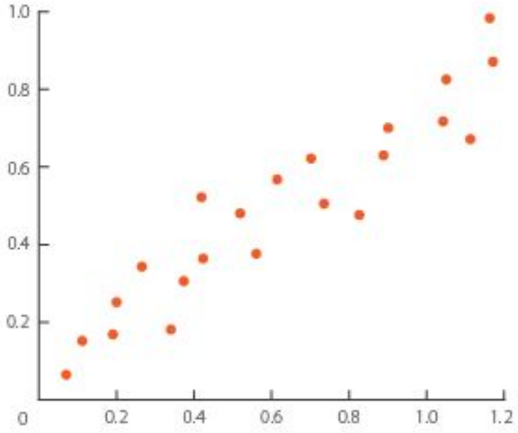
Types of data visualisation



datavizcatalogue.com/index.html

Scatterplot

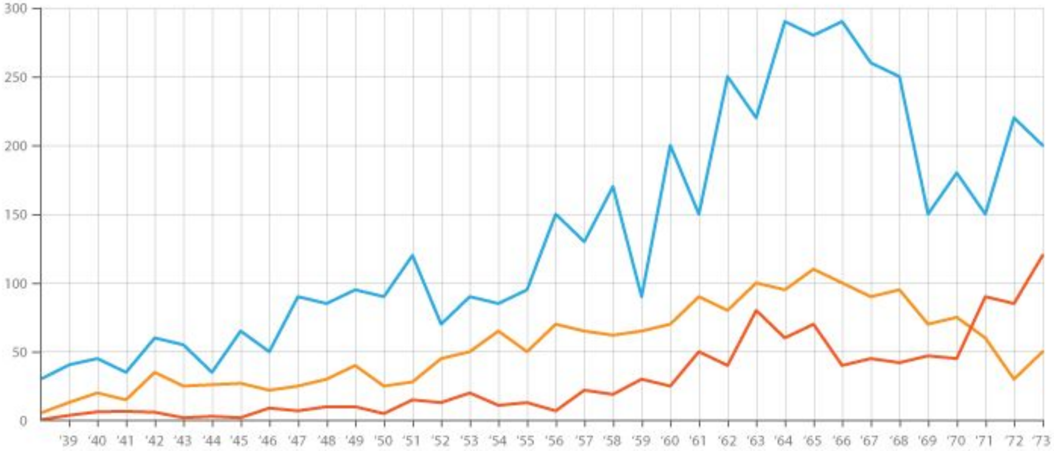
- Represent a collection of data points on an x y axis
- Show groups or correlations in the data
- The strength of the correlation is reflected in how densely packed the points are



datavizcatalogue.com/methods/scatterplot.html

Line

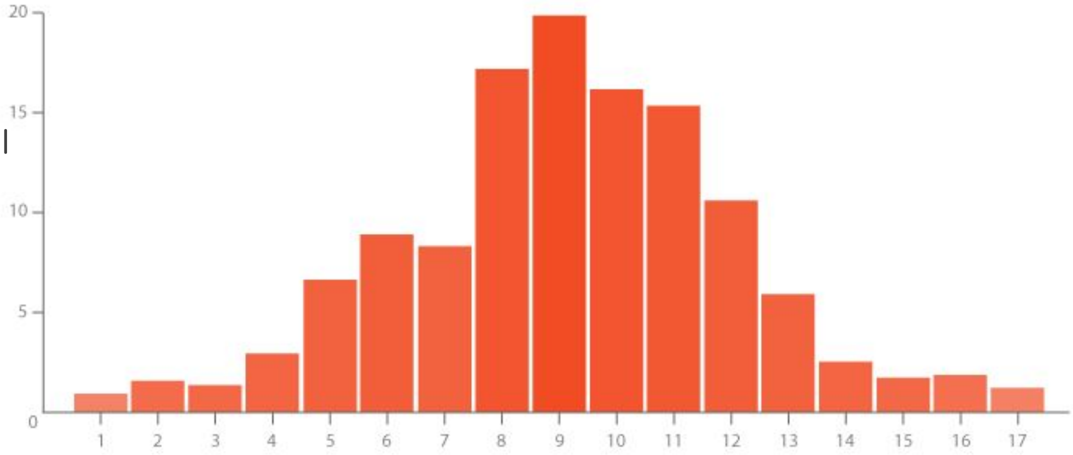
- Used to display data along a continuous scale
- Most often used to show change over time
- Avoid too many lines on a single graph



datavizcatalogue.com/methods/line_graph.html

Histogram

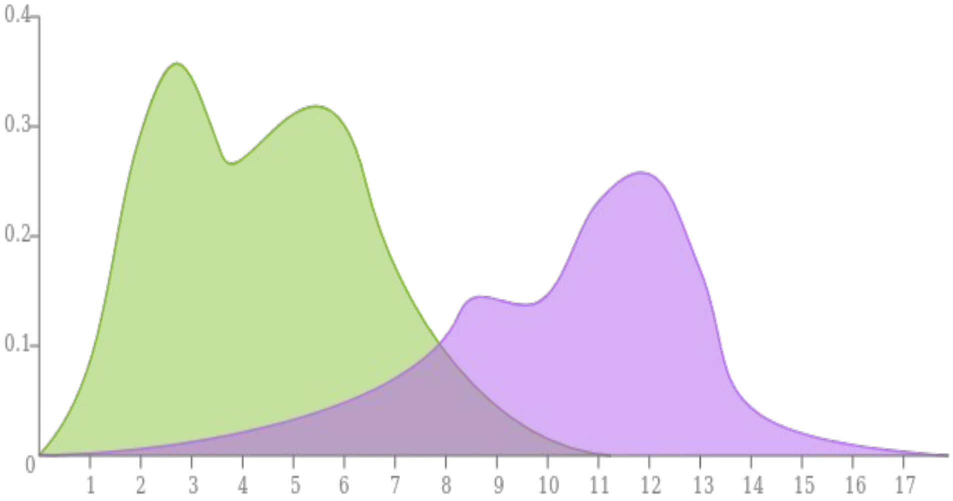
- Visualises the distribution of data over a continuous interval or time period
- Can be used to represent categorical data



datavizcatalogue.com/methods/histogram.html

Density

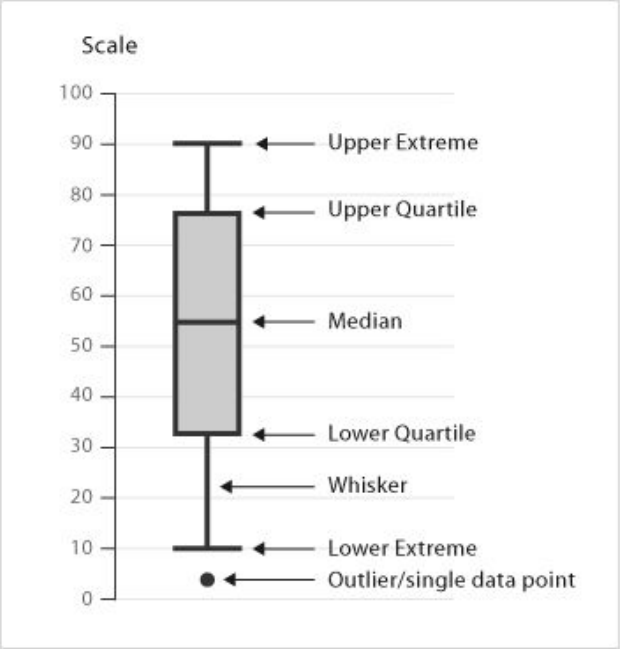
- Shows the distribution of data over a continuous interval or time period
- Gives greater detail than a histogram as it isn't affected by the number of bins



datavizcatalogue.com/methods/density_plot.html

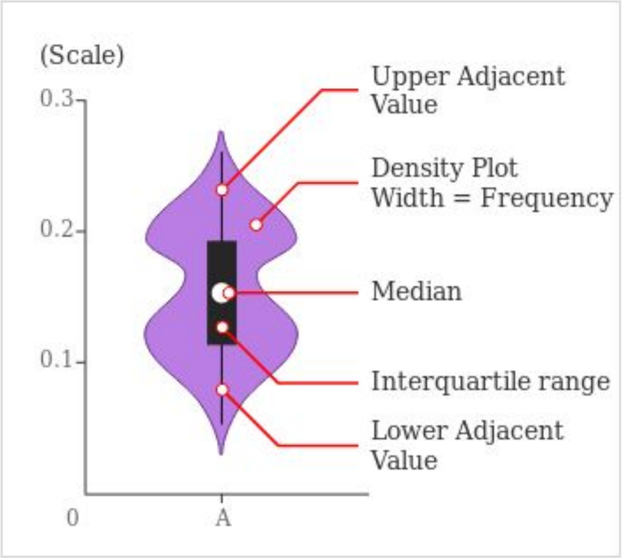
Box and whisker plot

- Displays key values: median, 25th percentile, upper and lower extremes
- Shows whether the data is symmetrical
- Shows how tightly is the data grouped and whether it is skewed
- Good for exploring large datasets



Violin plot

- Similar to a box plot, displays mean, interquartile range and distribution
- Width indicates frequency of a value
- Suitable for large amounts of data



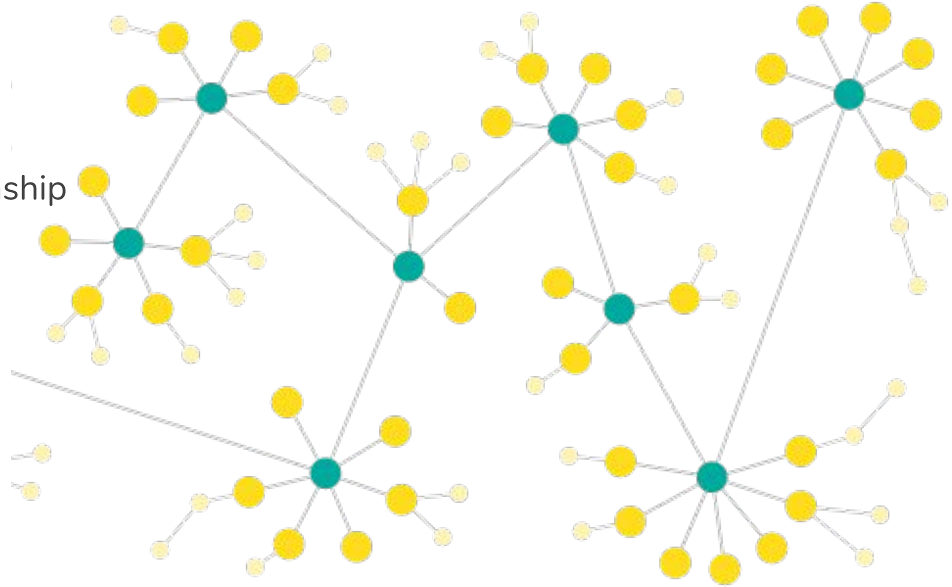
datavizcatalogue.com/methods/violin_plot.html

Pie

- Show proportion of a whole
- Frequently abused but do have a place

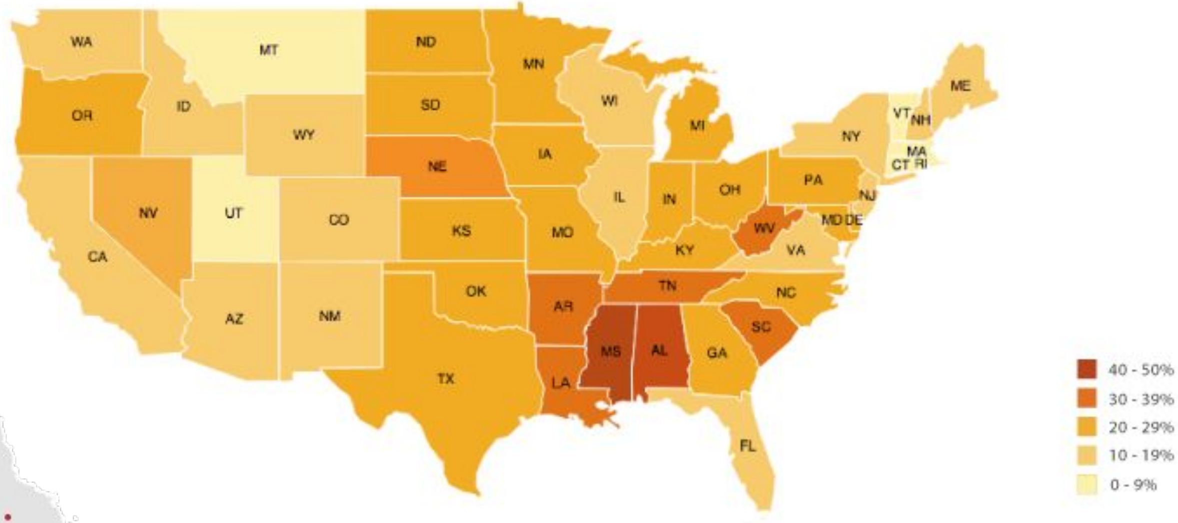
Network visualisation

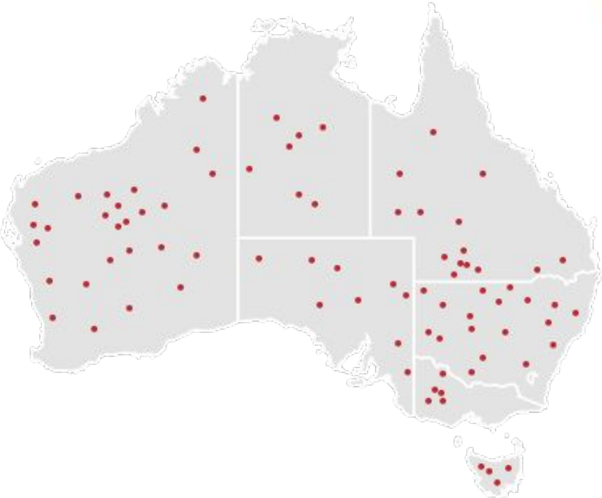
- Entities are dots or nodes
- Relationships are edges
- May depict the direction of a relationship
- Beware of hairballs!



datavizcatalogue.com/methods/network_diagram.html

Maps



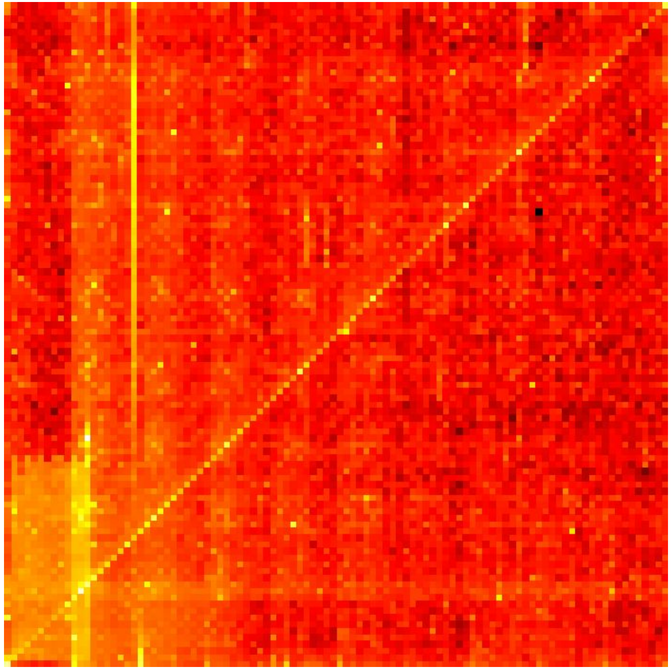


- Choropleth (top): regions are shaded to indicate presence of a variable
- Point (bottom): specific points are indicated, showing distribution of a variable

datavizcatalogue.com/methods/choropleth.html
datavizcatalogue.com/methods/dot_map.html

Heatmap

- Useful for exploring multivariate data
- Show a generalised view
- Helpful for detecting patterns



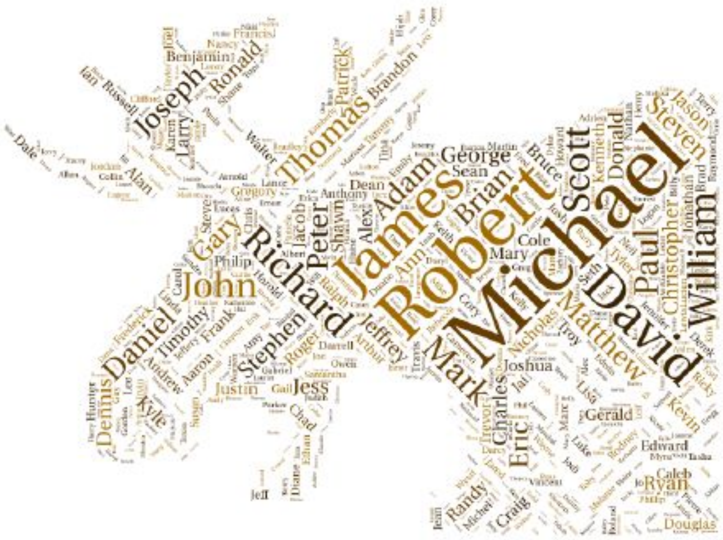
http://datagenetics.com/blog/september32012/grid.png

A word on wordclouds

- Long words may be overemphasised
- Need to stem words
- Not great for accuracy - more decorative

Names of Moose Hunters in Maine

2019 Maine Moose Permit Lottery Winners



Source Data: Maine Department of Inland Fisheries & Wildlife
Moose Image: Richard Lee, @brock222

Exercise: card sort

Intrinsic to each dataset is the best way to visualise it

Elements of design

As well as the type of visualisation you choose, you have a number of elements at your disposal to make your visualisation clear and effective

Size and scale

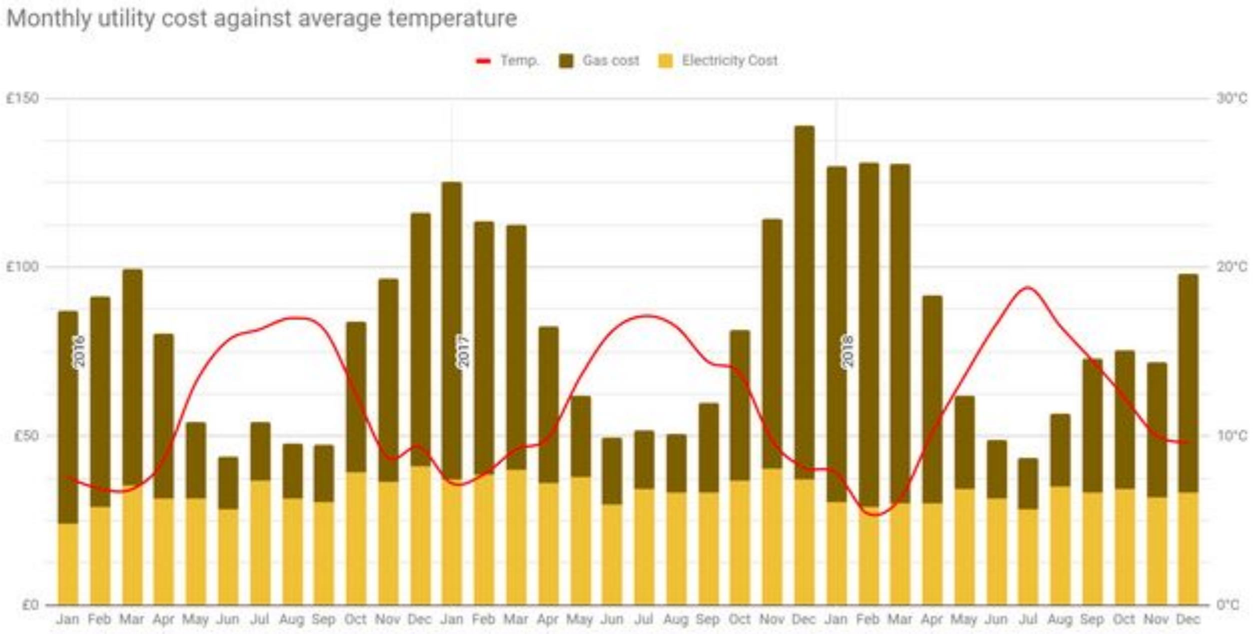
Colour

Labels

Angle

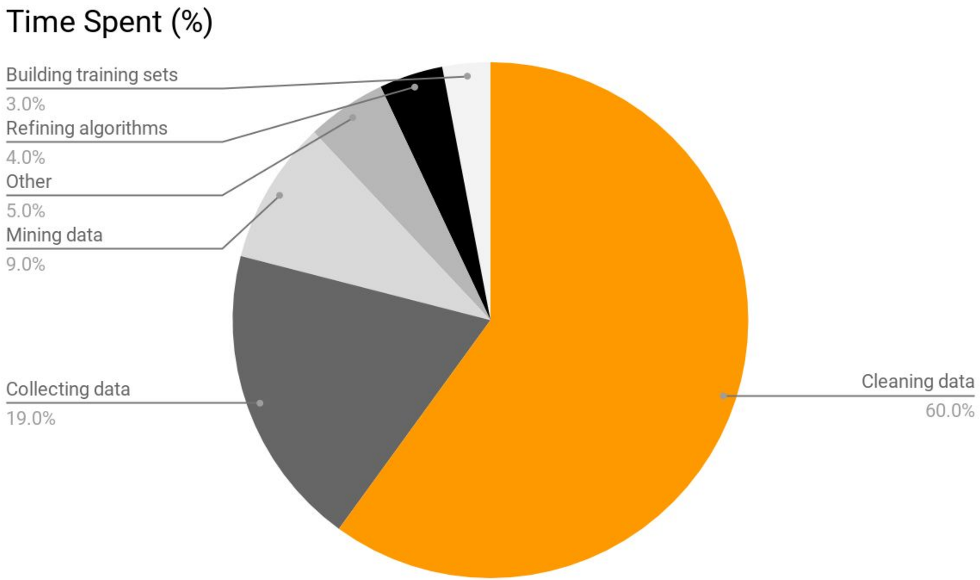
Grouping and selecting

Size and scale



reddit.com/r/dataisbeautiful/comments/bx839y/3yrs_of_my_monthly_utility_cost_against_avg/

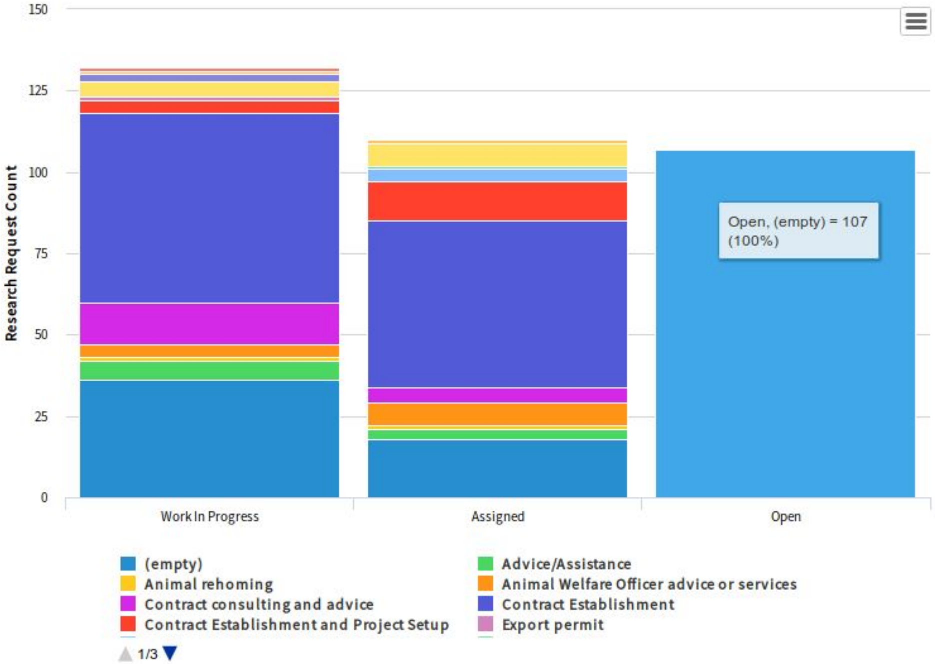
Colour: Highlight a key feature



Source: CrowdFlower 2016
whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data

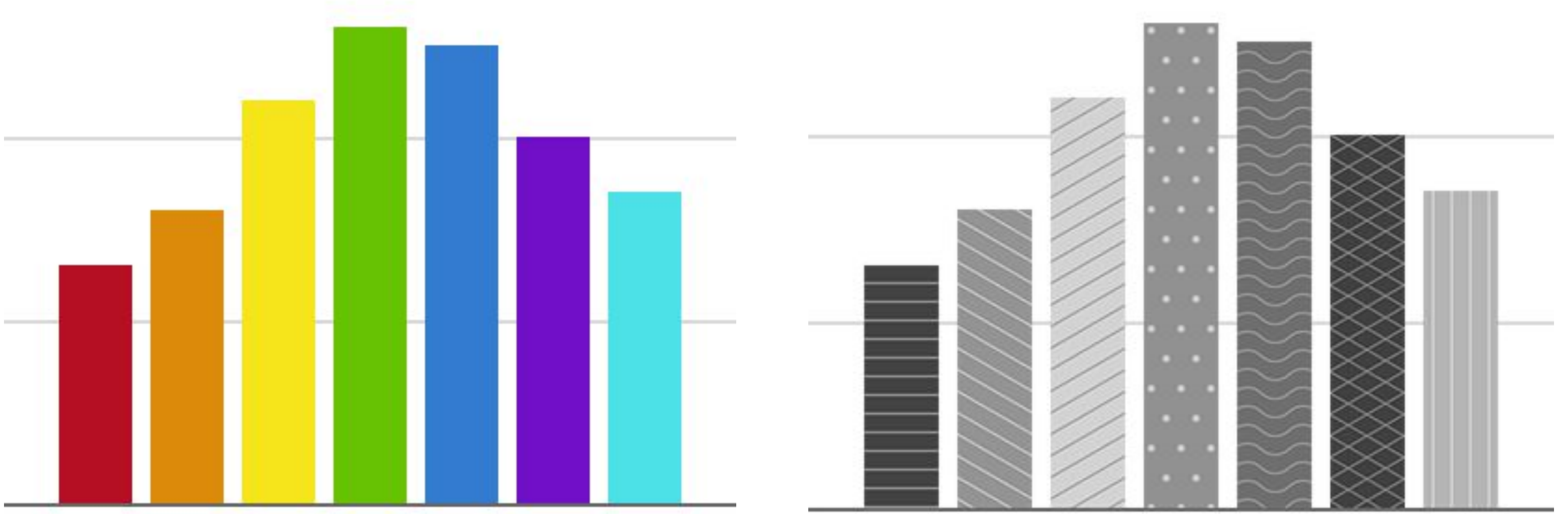
Caution: Unicorn Vomit

Tip - If your tool permits you to control Hue, Saturation and Lightness, vary only one of these to create a fairly harmonious colour palette



Accessibility

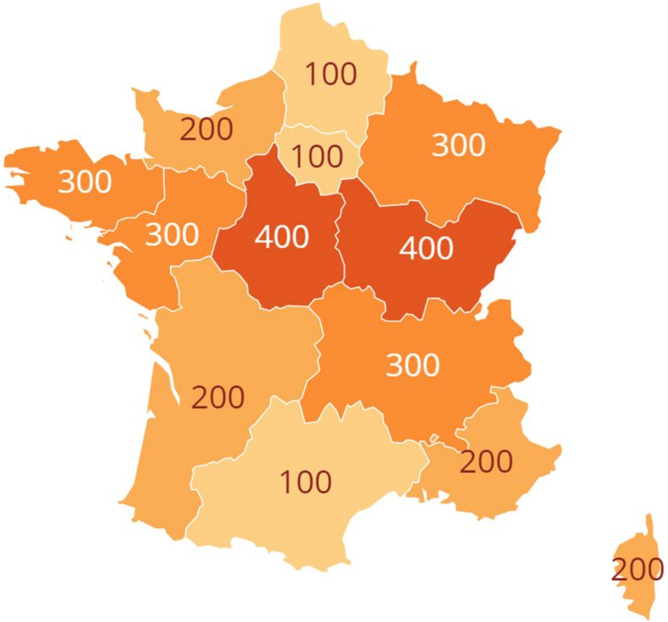
Colour-blind friendly options



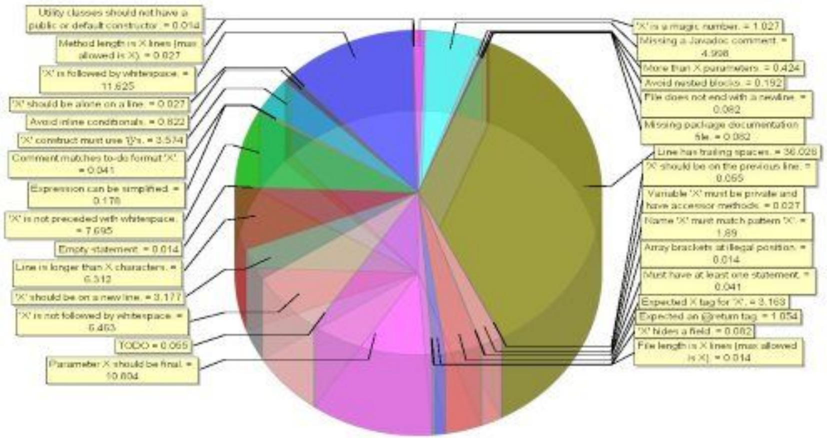
https://uxcellence.com/2018/accessible-color-contrast

Labels

Add clarity, except when they don't



knowledge.domo.com/Visualize/Adding_Cards_to_Domo/KPI_Cards/Building_Each_Chart_Type/World_Map

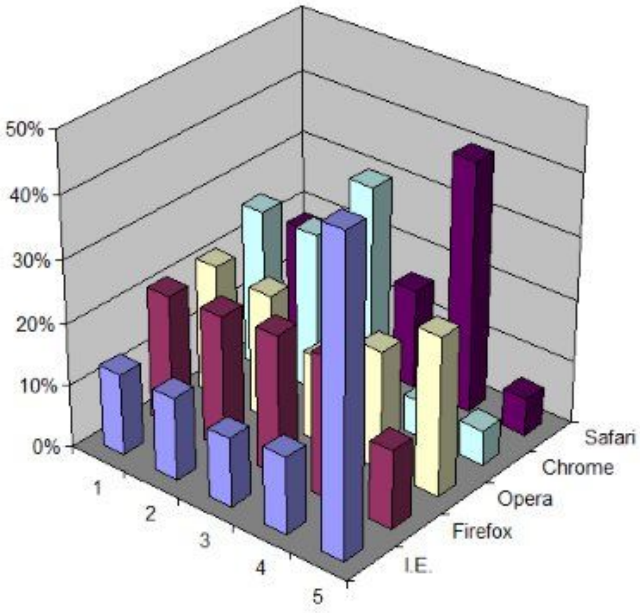


livingqlikview.com/the-9-worst-data-visualizations-ever-created

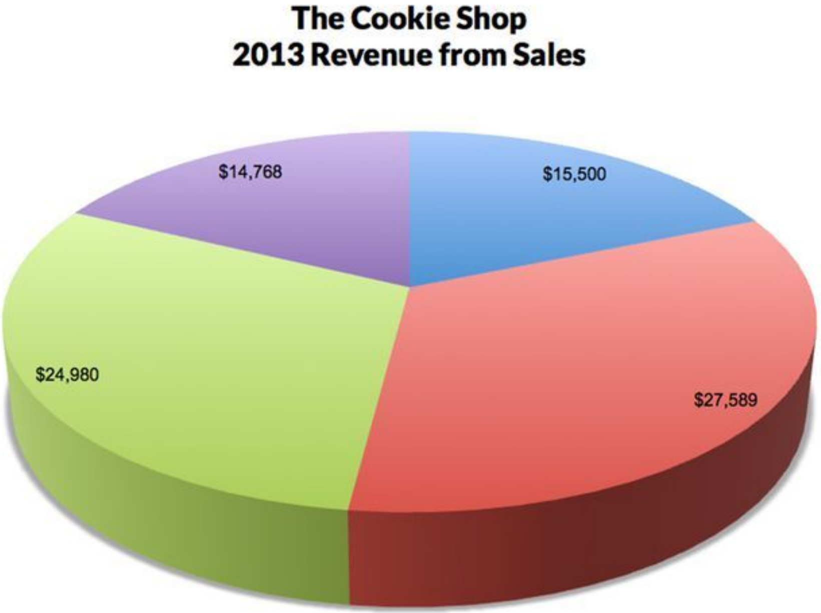
3D visualisations

Problematic as scale is distorted

Can’t see everything clearly



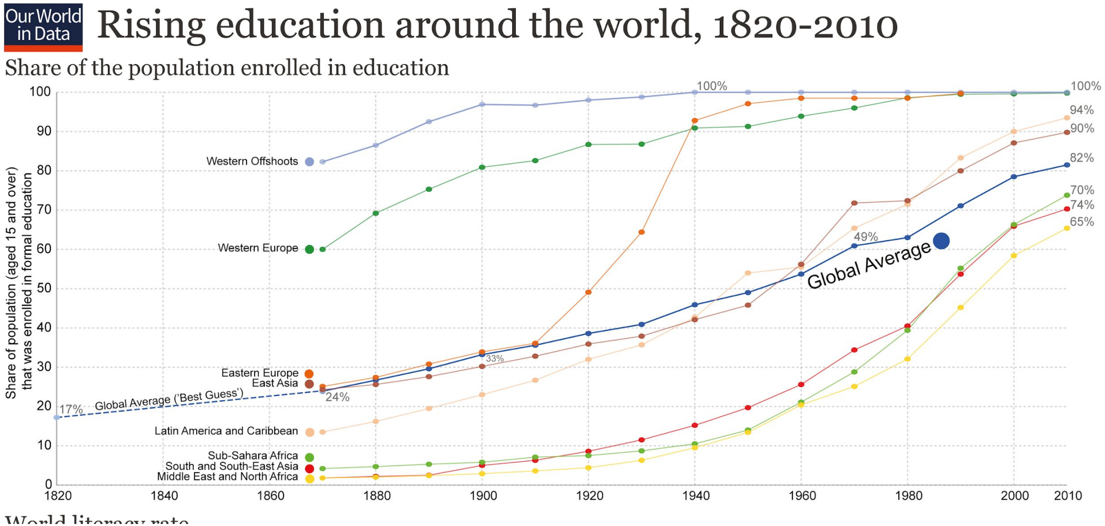
Creating distortion



https://www.lifewire.com/create-a-pie-chart-in-excel-3123565 ourworldindata.org/uploads/2018/09/Rising-Education-Around-the-World-School-and-Literacy.png

Grouping and selecting

- Grouping can make it easier to represent data with a lot of categories by reducing the number
- It can be used effectively to highlight a key statistic
- Hint: If your biggest group is ‘other’, you need to rethink your groupings


Our World in Data

Rising education around the world, 1820-2010

Share of the population enrolled in education

World literacy rate

Exercise: data is ugly

Find a terrible data visualisation - try reddit.com/r/dataisugly or viz.wtf for some great(?) examples

What’s wrong with it?

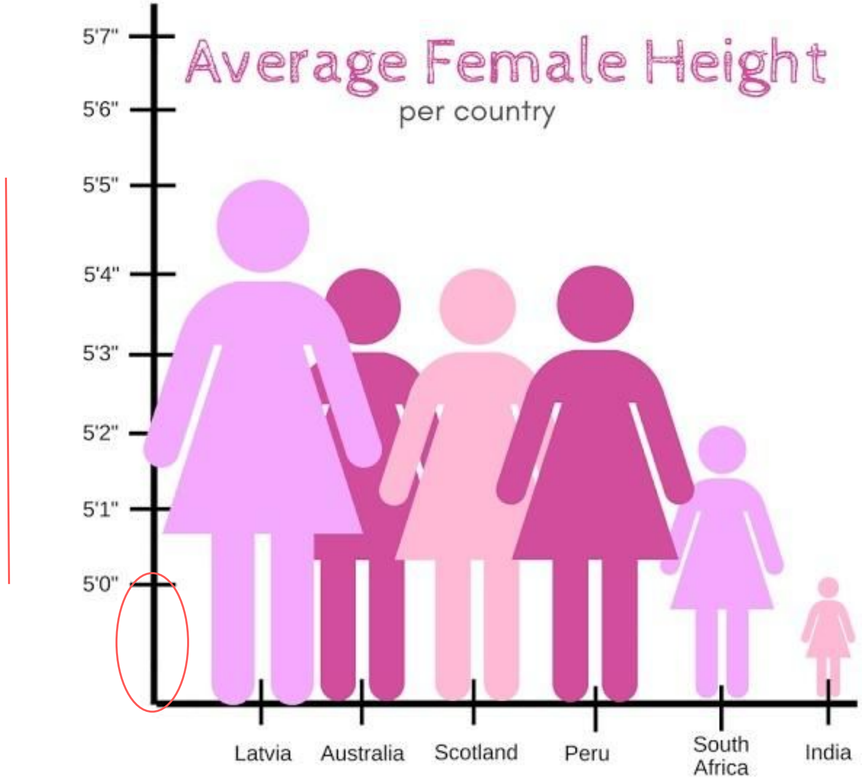
Share your example with your group

Pick the most interestingly terrible

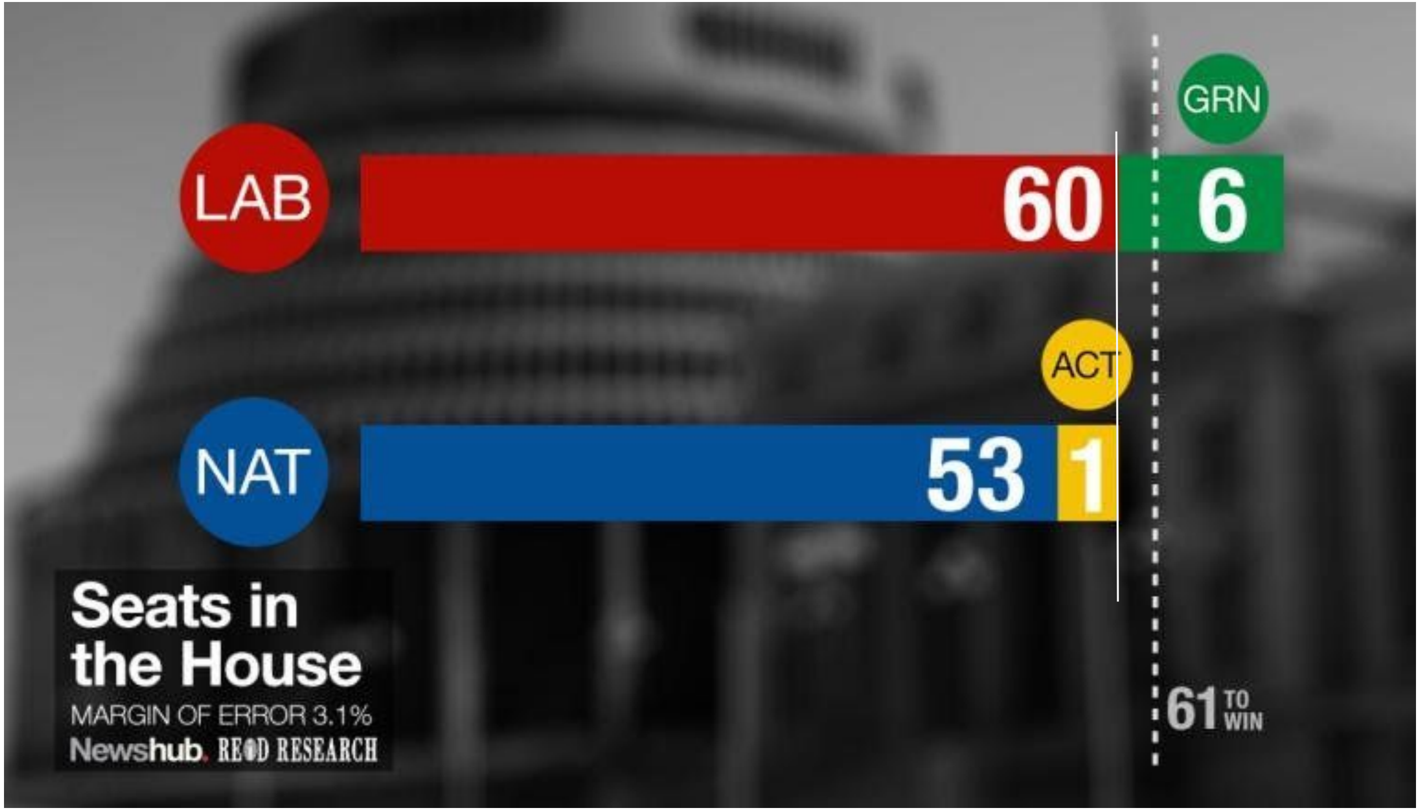
Share the link to the etherpad https://etherpad.net/p/MBS_DataVis

Lying with graphs

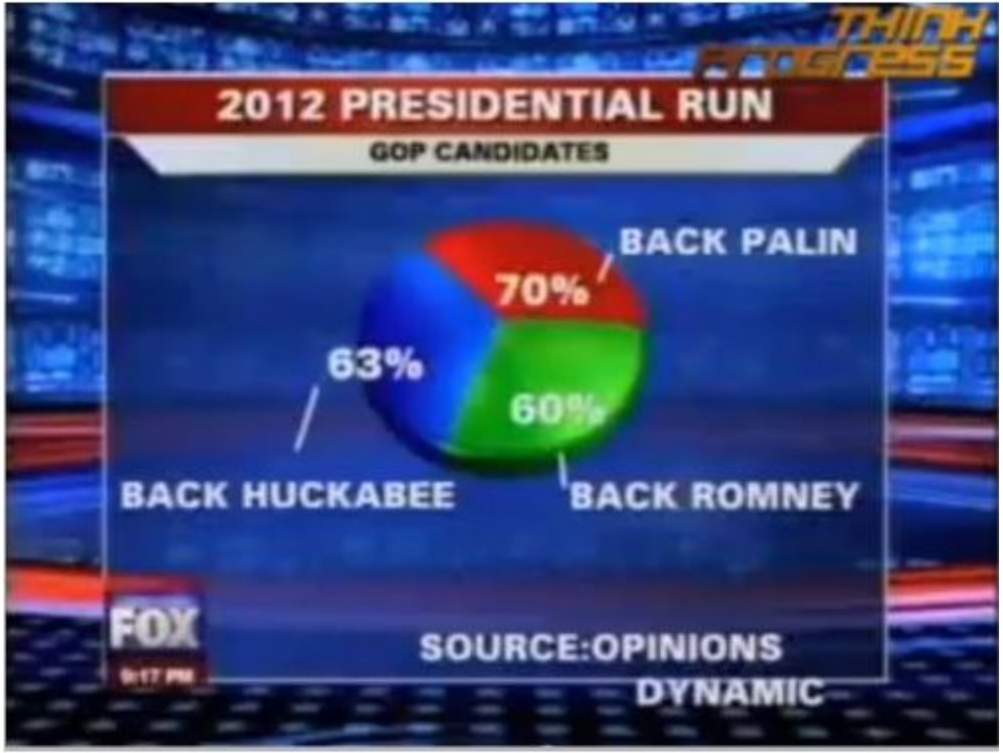
More than bad design…



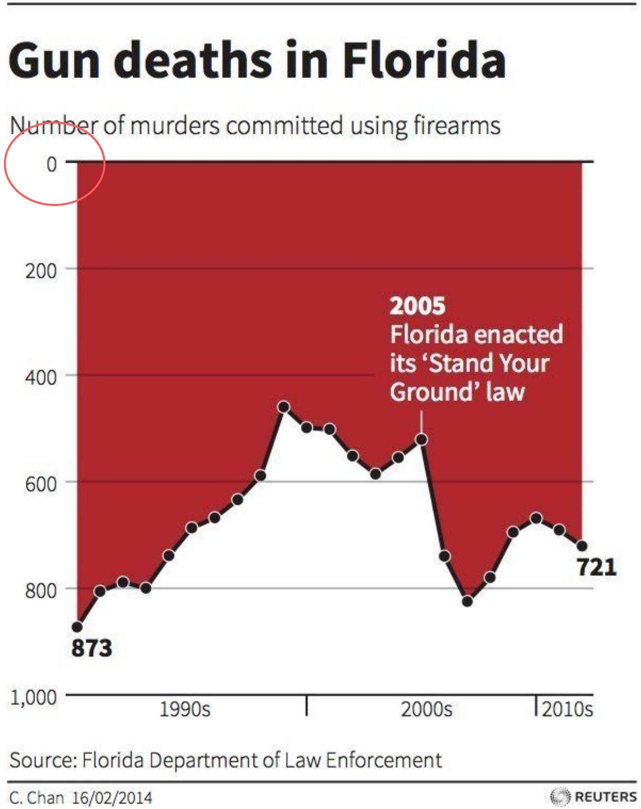
https://twitter.com/lizardbill/status/1127005323636686848

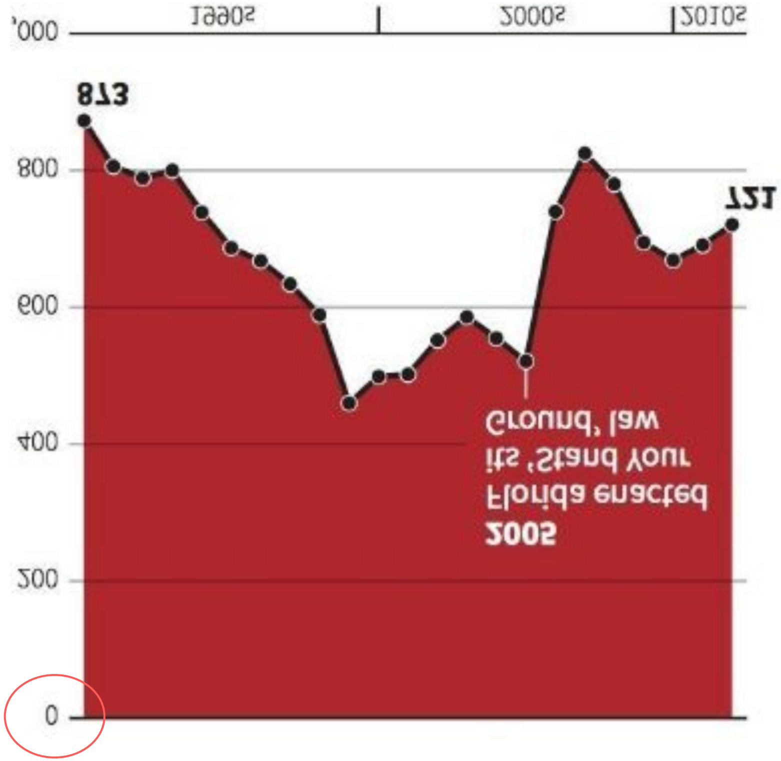


Distorted scale
https://twitter.com/marcdaalder/status/1094836212773179392



Further crimes against pie-charts
livingqlikview.com/the-9-worst-data-visualizations-ever-created





Inverted axis

Getting on with it

Your Dataviz workflow

- What is your question?
- Get data
- Inspect data
- What visualisation types are appropriate?
- Who is your audience?
- How will your visualisation be displayed?
- Prep data - record transformations
- **Analyse and visualise**
- Store data
- Store code
- Export and share

Data Science 101

Take a copy of the data available from the link on the etherpad
https://etherpad.net/p/MBS_DataVis

1. How large was each class?
2. How does confidence in programming compare to confidence in communication?
3. How confident is this class in statistics?
4. In this class, how does confidence in maths compare to confidence in business acumen?

Exercise: Offscreen data visualisation