Analyzing and Visualizing Survey Questions Using Open-Source Software

Okan Bulut

Centre for Research in Applied Measurement and Evaluation
University of Alberta

To download this presentation: http://bit.ly/otessa2022dataviz

Outline

01 Visualizing survey items

What are the key principles in developing visualizations?

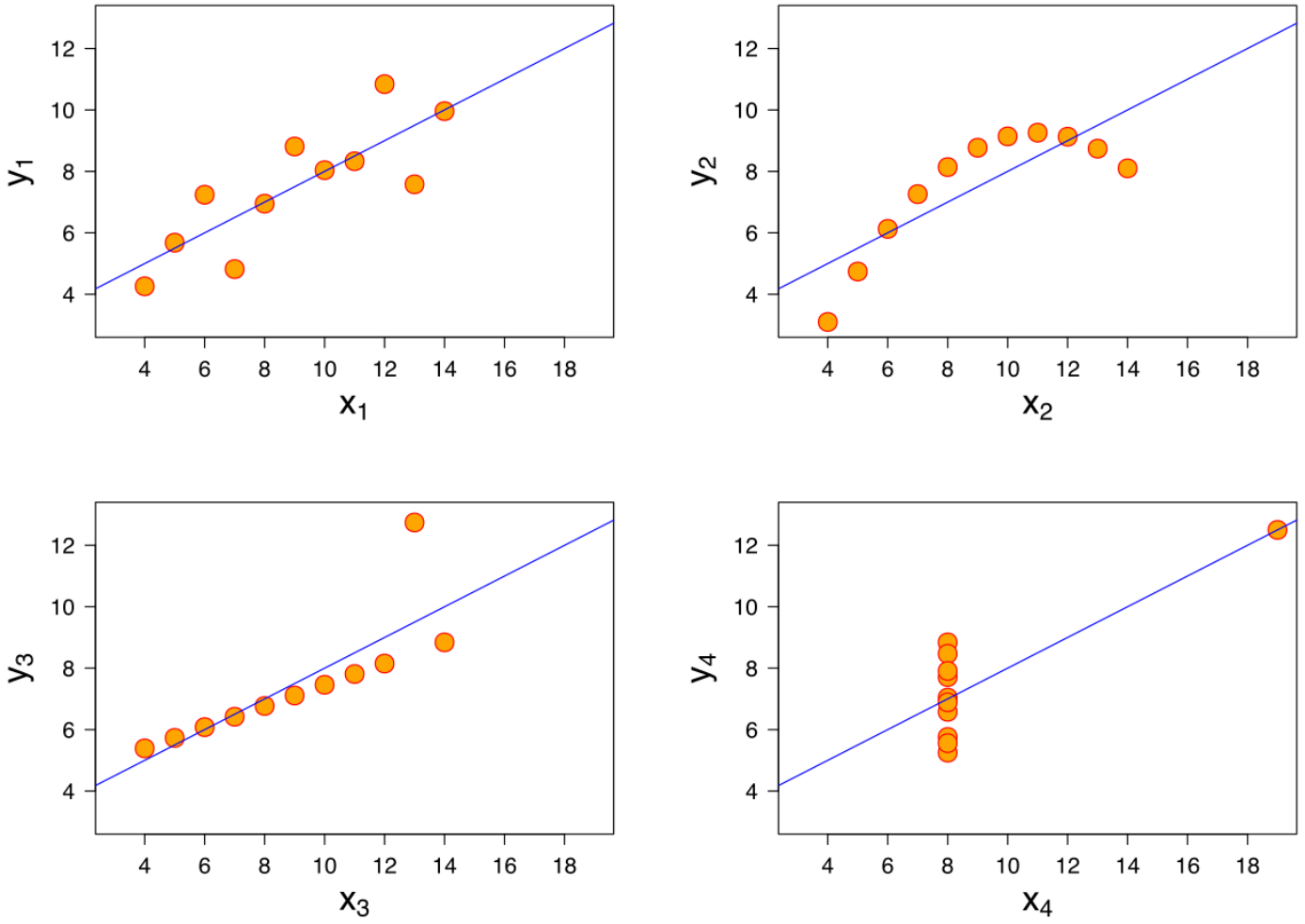
02 Evaluating survey items

What are the visual and statistical analysis options for evaluating survey items?

03 Example

Visual and statistical analysis of survey items in PISA 2015

Visualizing Survey Items



Four datasets with nearly identical simple descriptive statistics for x and y but they have very different distributions…

Property	Value
Mean of x	9
Mean of y	7.50
SD of x	3.32
SD of y	2.03
Correlation of x and y	0.82

Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

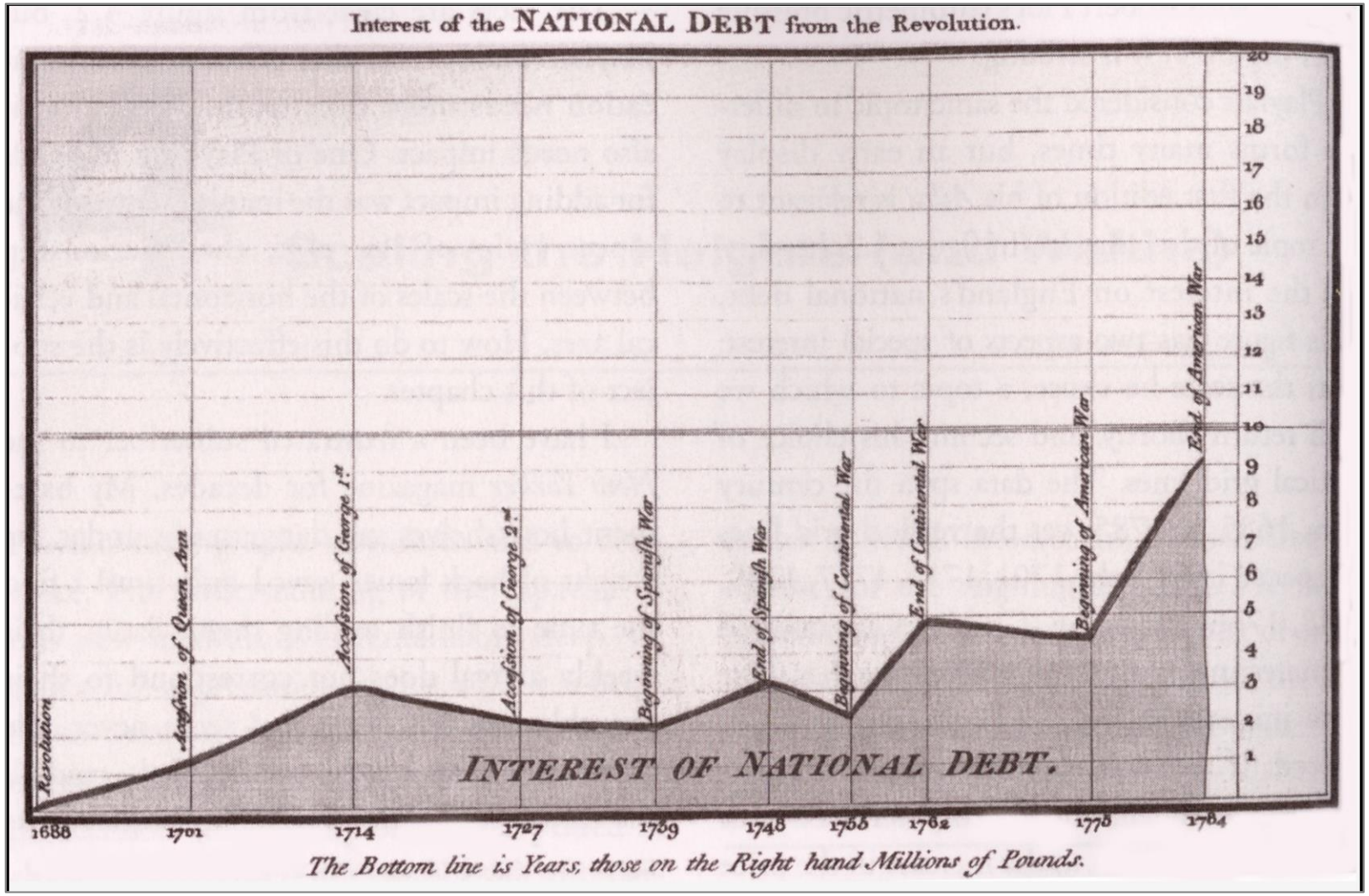
To move a huge amount of information into the brain very quickly

To identify patterns and communicate relationships and meaning

To inspire new questions and further exploration

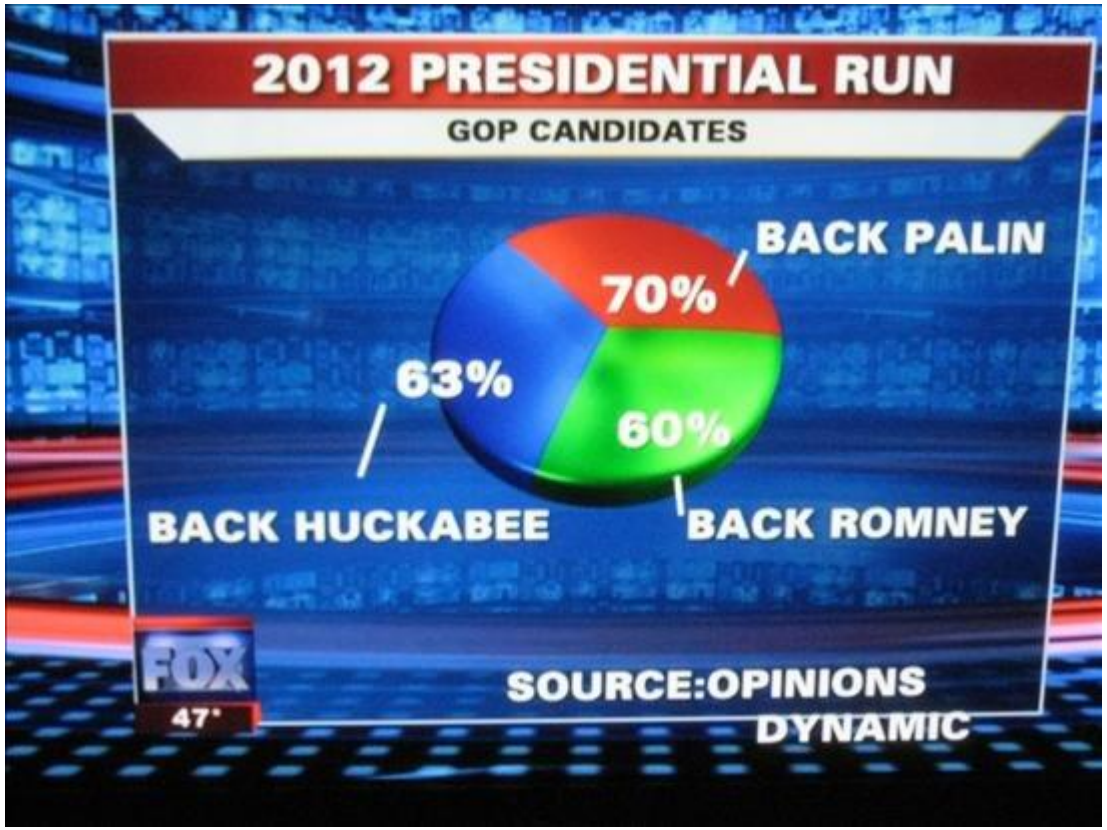
To help identify sub-problems

To discover or search for interesting or specific data points in a larger field



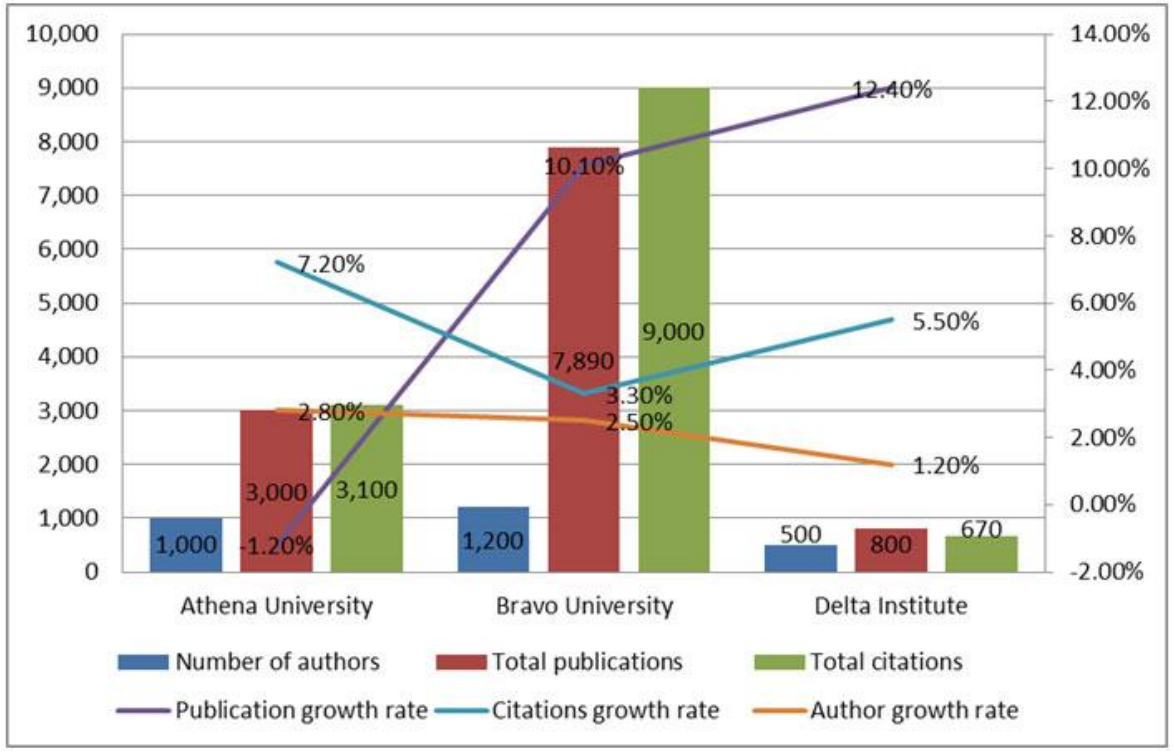
Source: Hand drawn by William Playfair (1786) in The Commercial and Political Atlas – *to make a case against England's policy of financing colonial wars through national debt*.

Did we get any better?

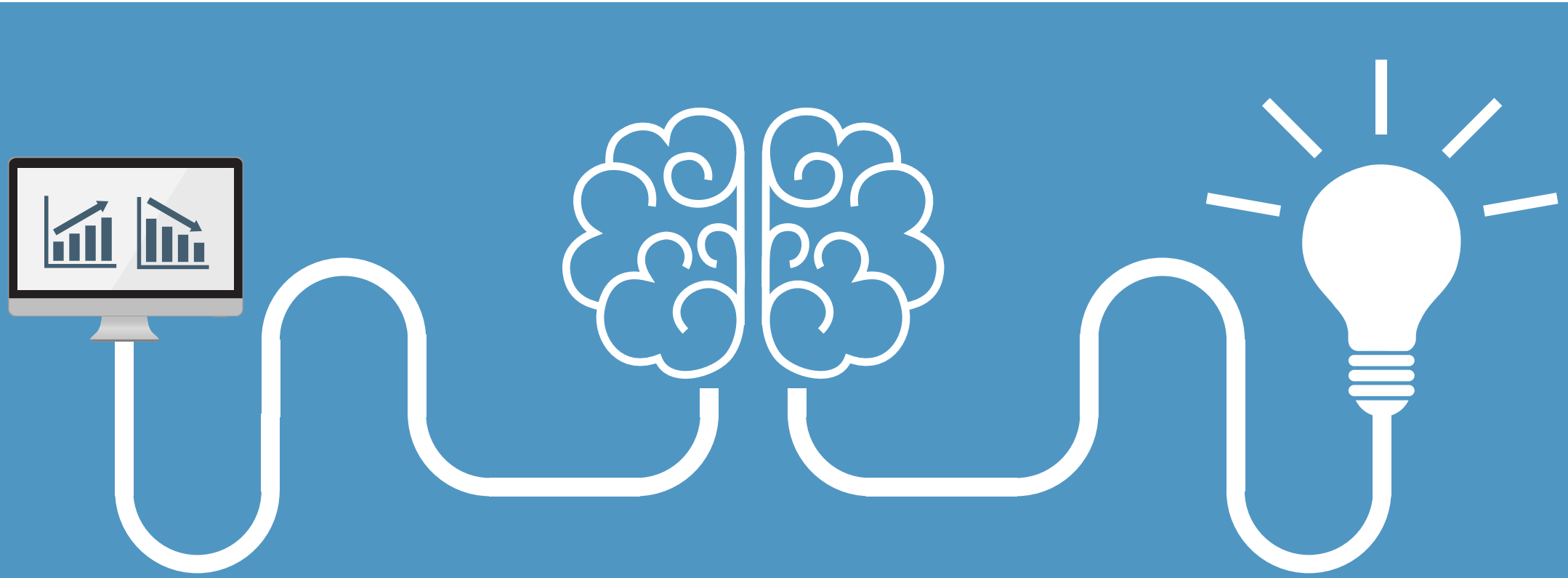


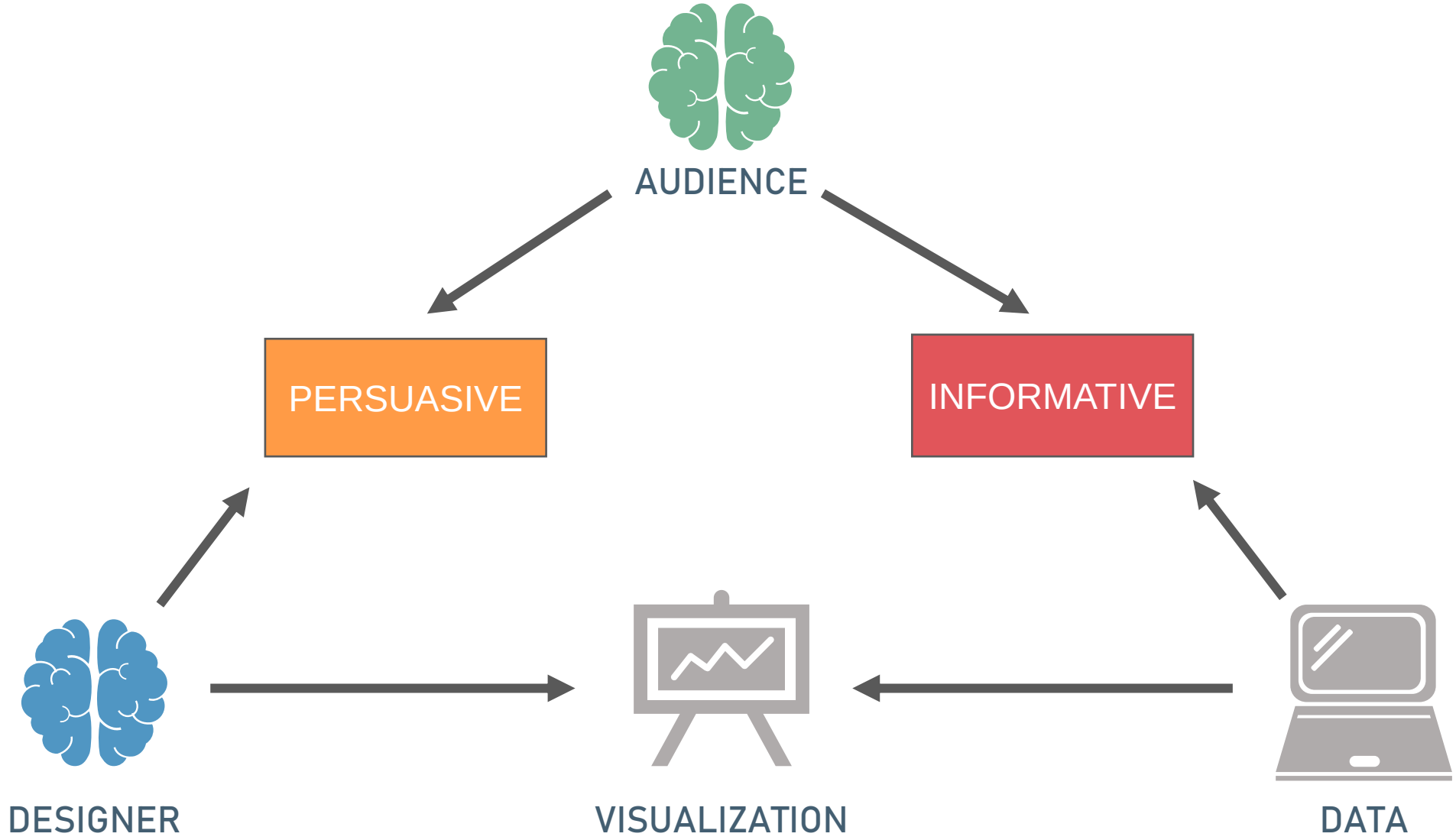
Source: Fox News – the percentages add up to 193%...

We have all done this…



Purpose





Source: Adapted from Iliinsky and Steele (2011, p. 9)

In practice, we…

EXPLORE (Informative)

Potential issues in the data:

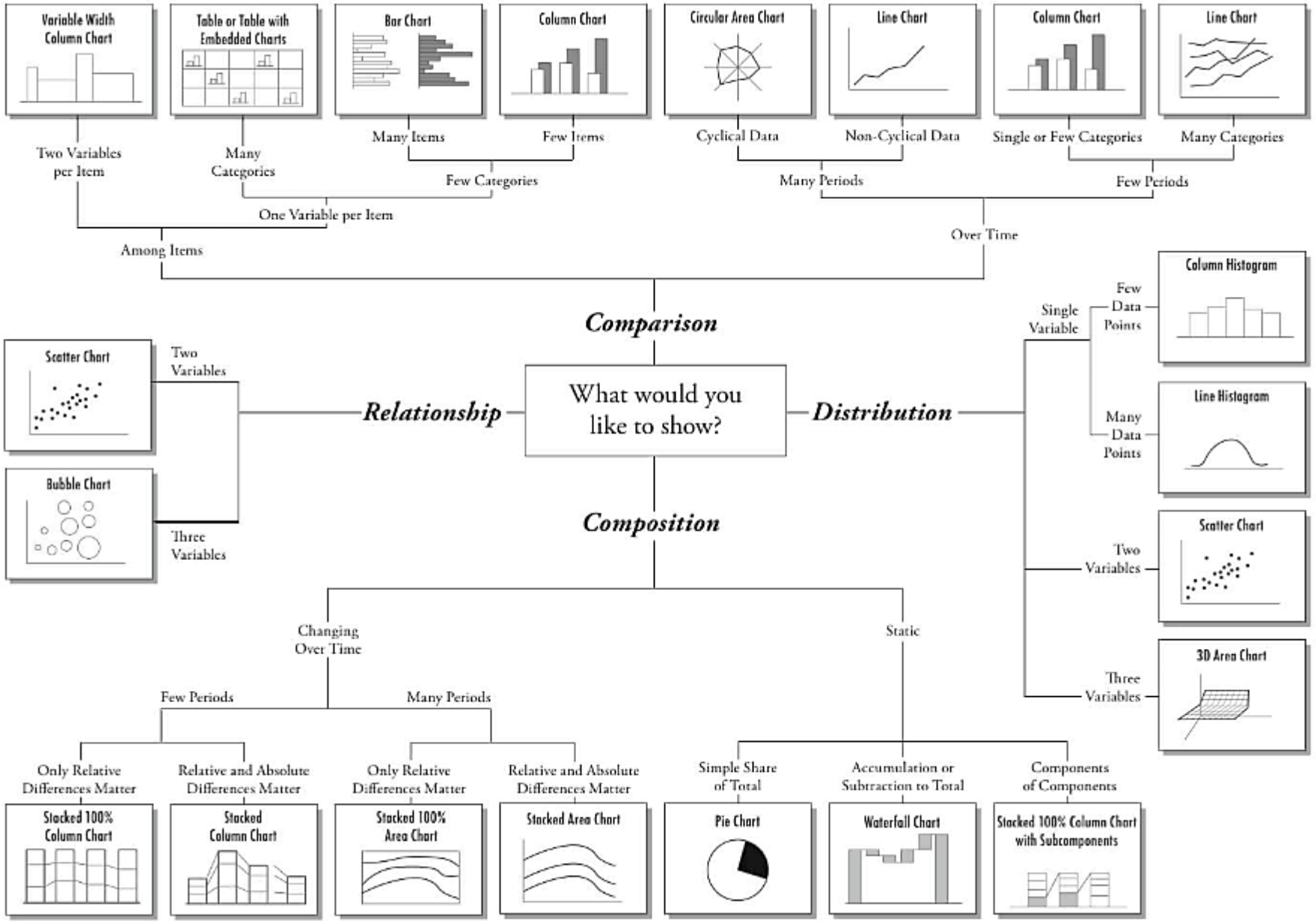
- Missingness
- Outliers
- Non-normality
- Non-linearity
- Extreme skewness and kurtosis

EXPLAIN (Informative)

Relationships between variables; correlations; interactions; patterns over time

PROVE (Persuasive)

Statistical models (e.g., regression); model fit; accuracy; predictions; inferences



Source: https://extremepresentation.com/

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 between 2 variables 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/or 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 long 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 parliaments – 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 cohesively-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

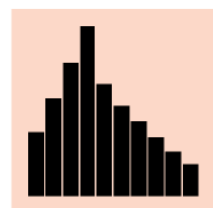
Used for showing the strength and inter-connectedness of relationships of varying types.

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.

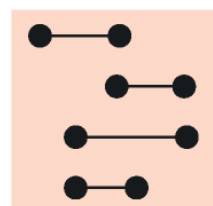
Example FT uses
Income distribution, population (age/sex) distribution, revealing inequality

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.

Determine the number of dimensions to be adjusted

Number of variables

Colours and shading

Shapes and lines

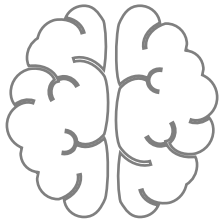
Size/area and line weight

Font type and font size

Decoding → Understanding

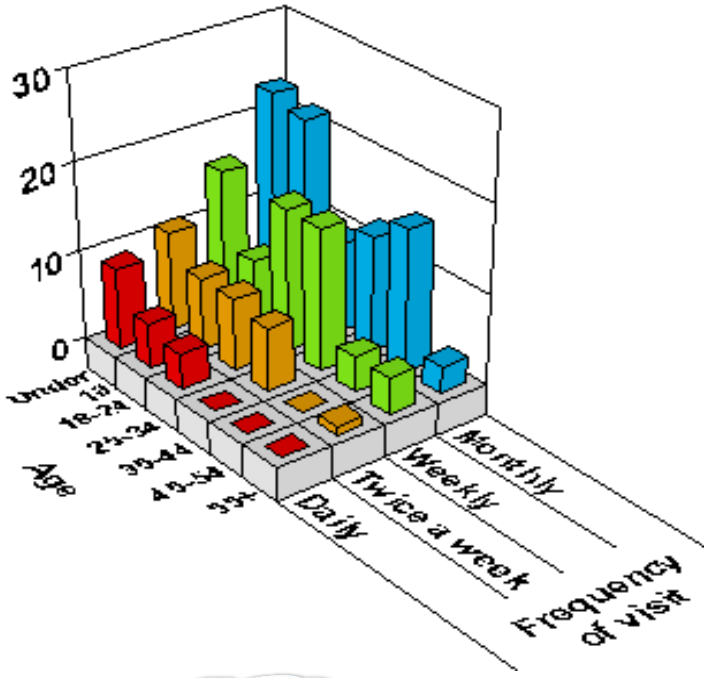
Brainpower used for **decoding**	Brainpower left for **understanding**

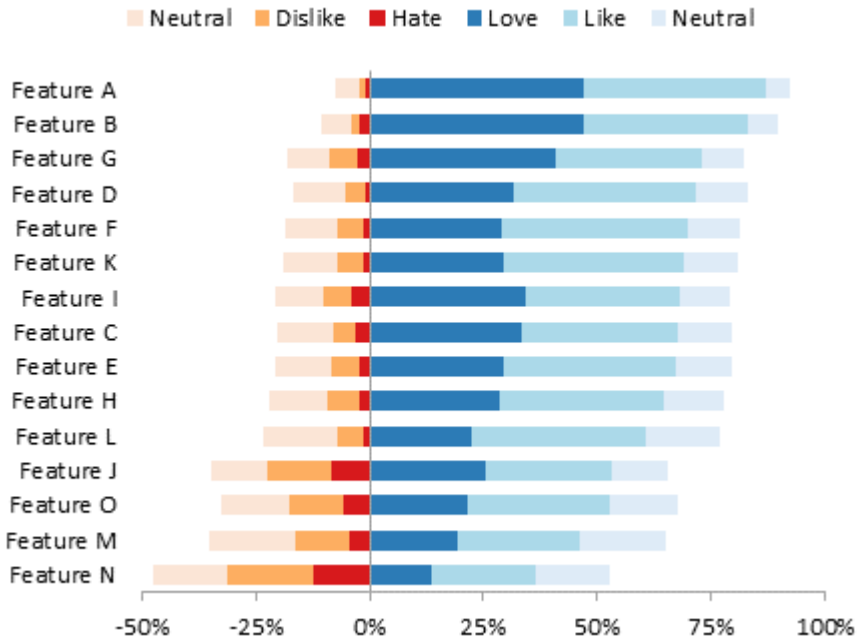
TOTAL BRAINPOWER AVAILABLE



Source: Adapted from Iliinsky and Steele (2011, p.24)

More Complex ≠ Better





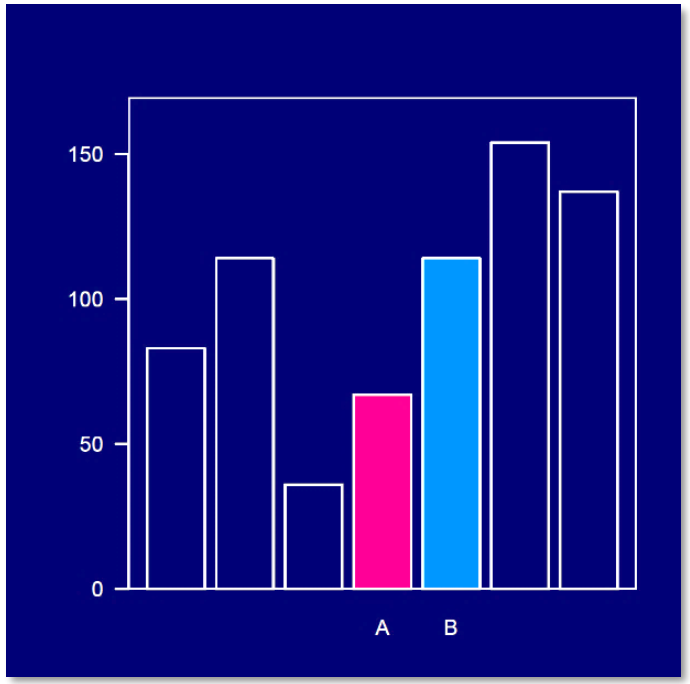
“**Simplicity** is the ultimate sophistication.”

-- Leonardo da Vinci

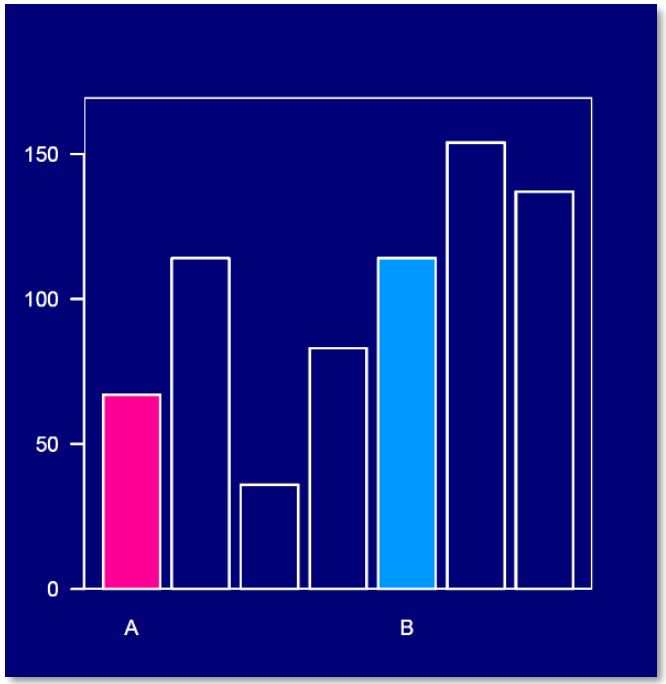
Source: https://www.darkhorseanalytics.com/blog/data-looks-better-naked

Source: https://www.darkhorseanalytics.com/blog/salvaging-the-pie

Which comparison is easier?

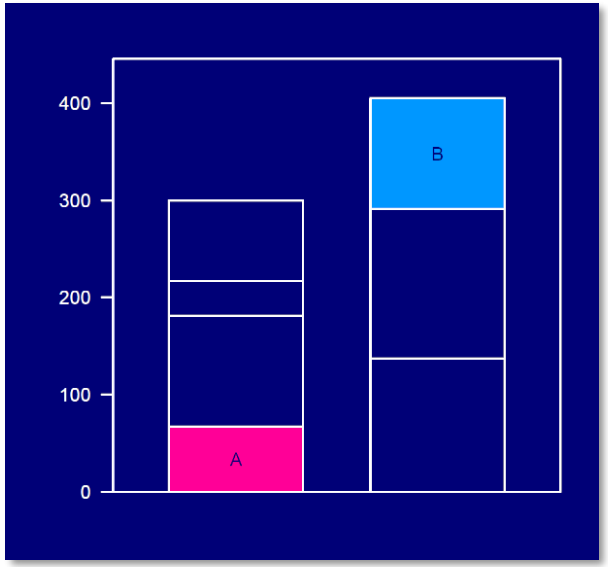


1

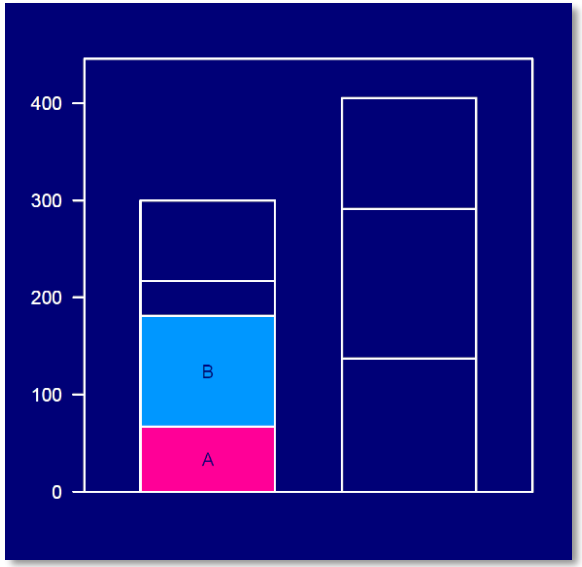


2

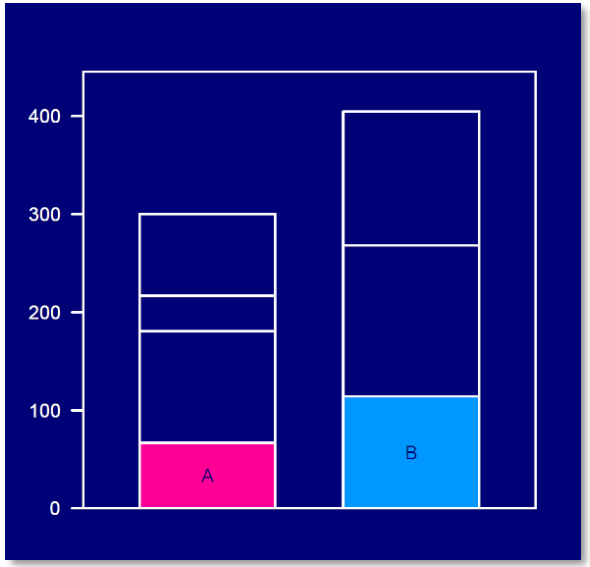
Which comparison is the easiest?



1

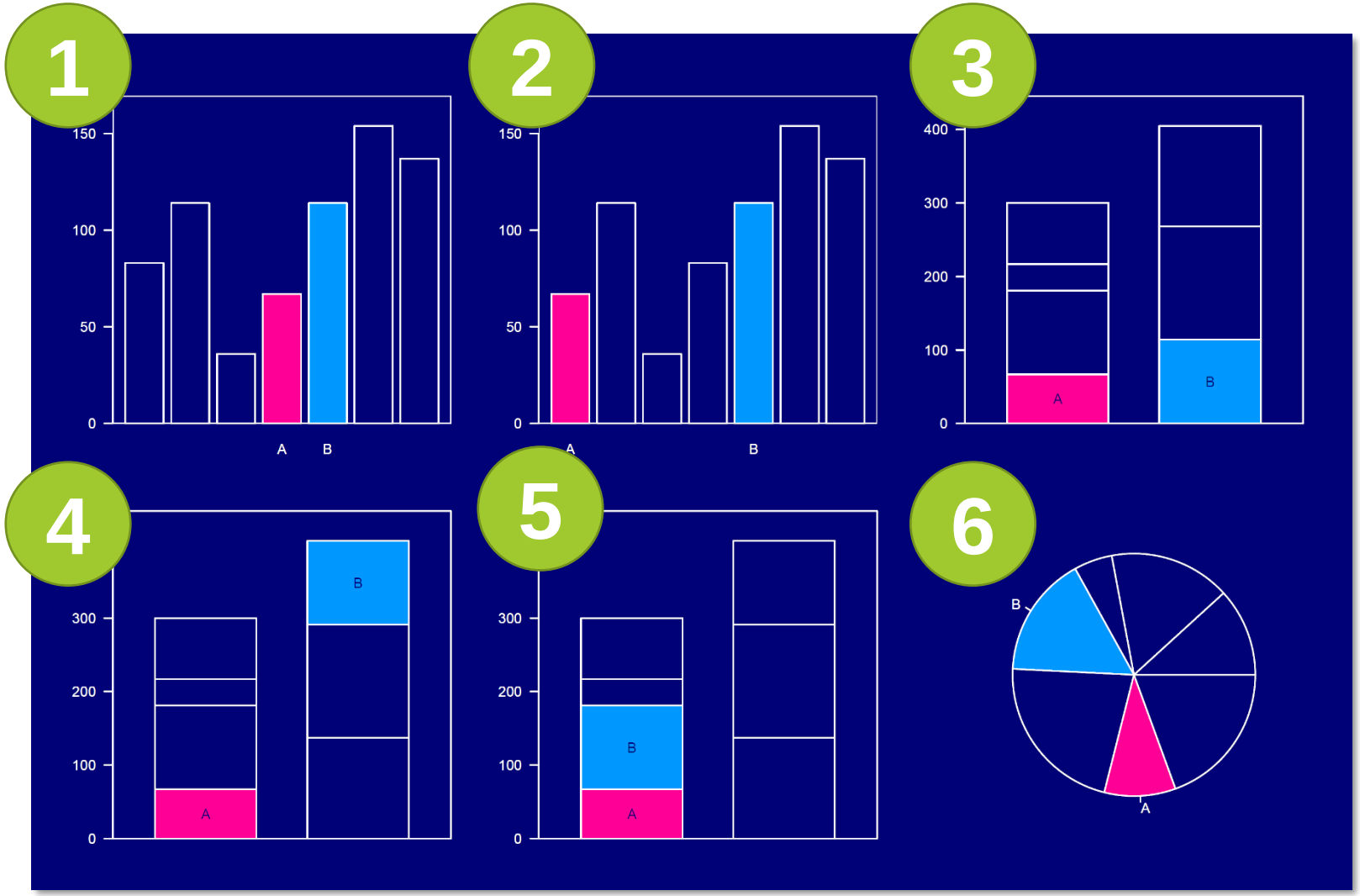


2



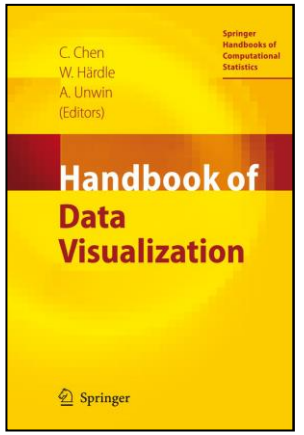
3

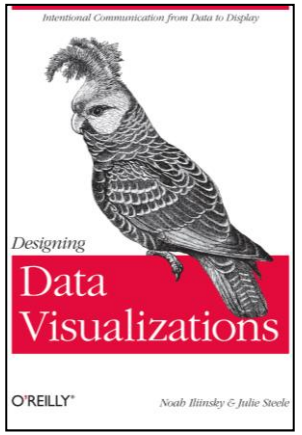
Which comparison is the easiest?

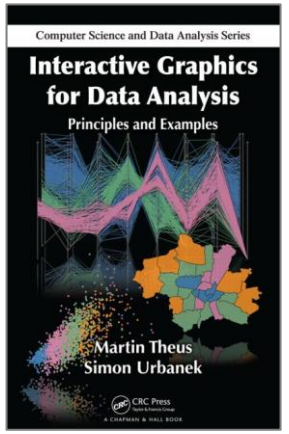


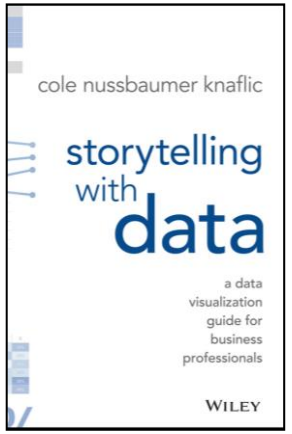
Some Resources…

- Stephanie Evergreen - **Data Visualization Checklist**
- Financial Times - **Chart Doctor**
- Darkhorse Analytics - **Visualizing Distributions**
- Chez Voila - **Glass Ceiling Visuals Remake**
- Eager Eyes - **Understanding Pie Charts**









https://towardsdatascience.com/5-ways-to-effectively-visualize-survey-data-using-r-89928bf08cb2

Data Visualization Software

Software Options (1)

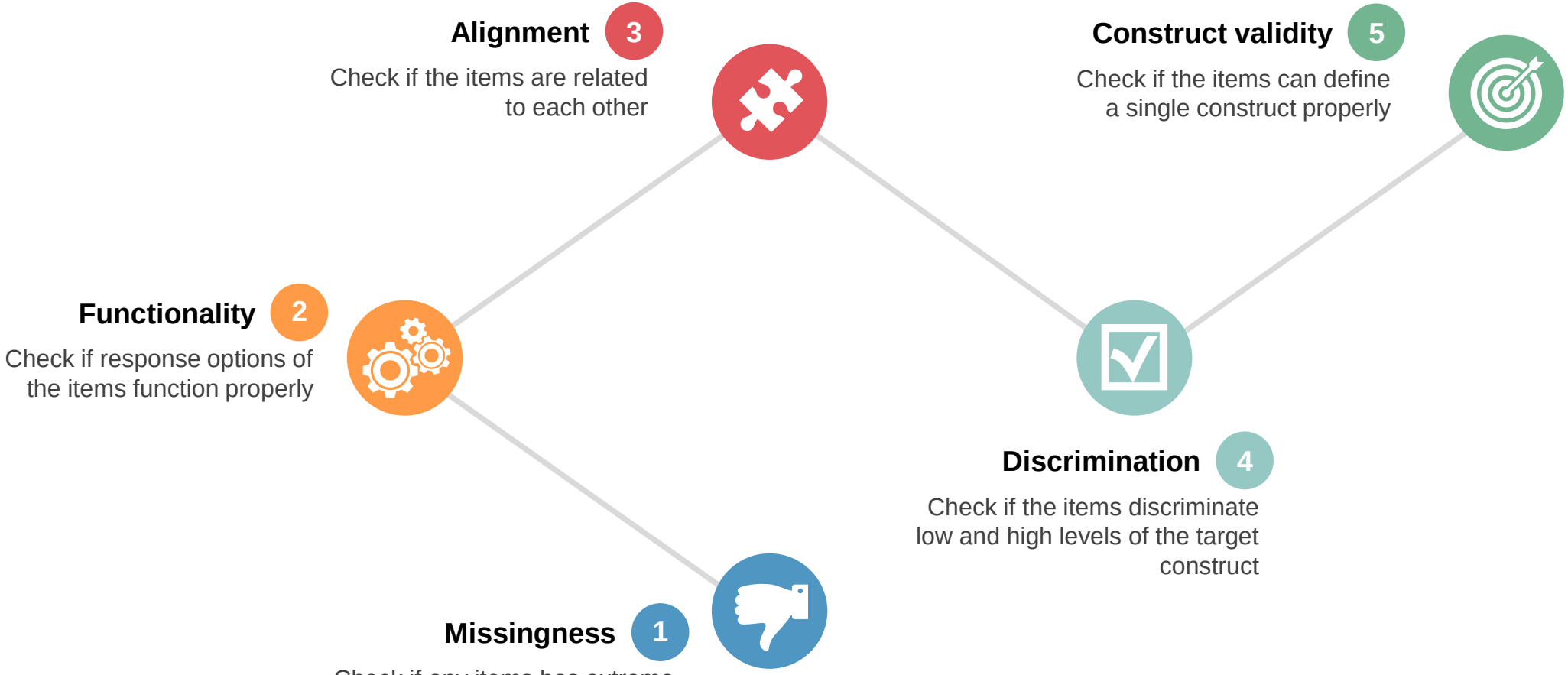
https://www.microsoft.com https://powerbi.microsoft.com/ https://www.alteryx.com/ https://plot.ly/ https://cran.r-project.org/

Software Options (2)

https://datastudio.google.com/ https://www.tableau.com https://www.datawrapper.de/ https://flourish.studio/ https://infogram.com/

Evaluating Survey Items

Checklist for Evaluating Items



Analyzing Survey Items

Descriptive statistics

For quantitative variables (e.g., age, height, weight):

- Mean, median, standard deviation, variance, minimum, maximum, etc.

For qualitative variables (e.g., Likert-scale items, demographic variables):

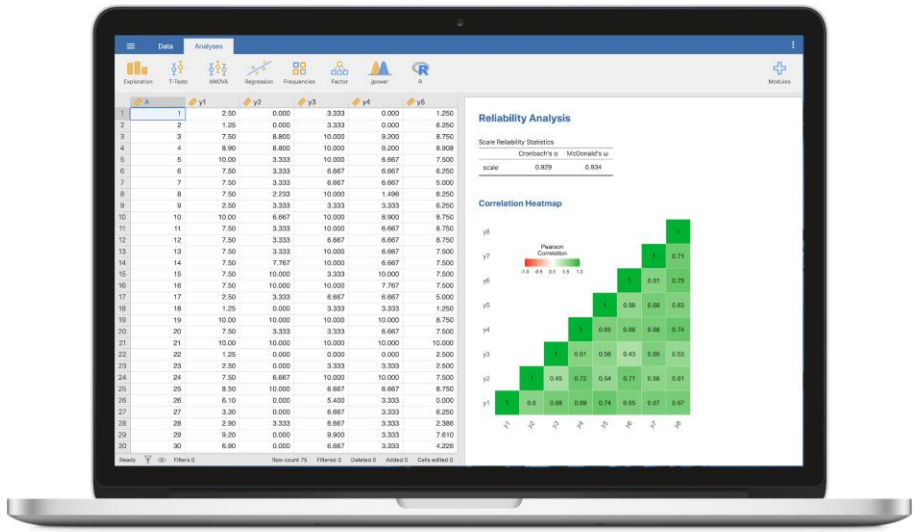
- Nominal variables: Frequencies and proportions (i.e., percentages)
- Ordinal variables: Frequencies, proportions, median, and mode

Categorical data analysis

Test of independence: This test examines whether two variables are either independent or related to each other (e.g., is there any relationship between gender and movie ratings (1 star, 2 stars, 3 stars, or 4 stars)?

Data Analysis Software

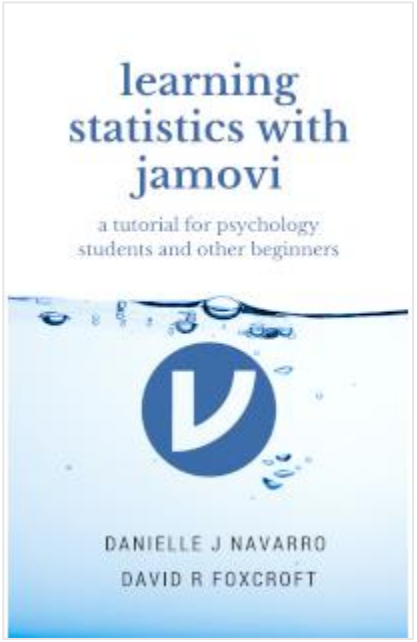
Software Options



jamovi
https://www.jamovi.org

JASP
https://jasp-stats.org

FREE (using R in the background)
Compatible with Windows, Mac, and Linux
Good for both statistical analysis and data visualizations



https://www.learnstatswithjamovi.com

https://datalab.cc/jamovi

jamovi Stats. Open. Now.

https://www.jamovi.org/user-manual.html
https://www.jamovi.org/community.html

Example

PISA

- http://www.oecd.org/pisa/
- A large-scale, international assessment for 15-year-old students
- Administered every 3 years
- 540,000 students from 72 countries participated in PISA 2015
- Reading, science, and math assessments (plus additional subject areas)
- Student, teacher, and school survey items to learn more about students

PISA

- Alberta students who participated in PISA 2015 (*n* = 2,133)
- Data files are available at: https://github.com/okanbulut/otessa2022
  - PISA_Alberta.csv
  - PISA_Alberta.sav

To run data analysis with jamovi, your dataset must be in one of the following formats:
✓ Excel (.xlsx)
✓ CSV (.csv)
✓ SPSS (.sav)
✓ R (.RData) or SAS (.xpt, sas7bdat)

- 10 Likert-type survey items **potentially** measuring “attitudes towards teamwork”
- Each question has the following response options:

1 = Strongly disagree **2** = Disagree **3** = Agree **4 = Strongly agree** **9999 = Missing**

First eight questions share the same statement:
"To what extent do you disagree or agree about yourself?"

1. I prefer working as part of a team to working alone.
2. I am a good listener.
3. I enjoy seeing my classmates be successful.
4. I take into account what others are interested in.
5. I find that teams make better decisions than individuals.
6. I enjoy considering different perspectives.
7. I find that teamwork raises my own efficiency.
8. I enjoy cooperating with peers.

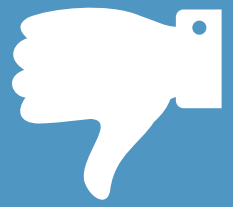
The other two items are independent:

9. I make friends easily at school.
10. Other students seem to like me.

Additional variables:

- studentid
- grade
- age
- gender (1 = Female, 2 = Male)

Missingness



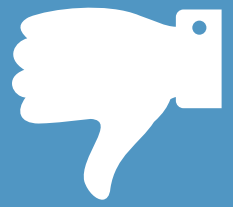
1. Import the data into jamovi.
2. Exploration → Descriptives

Descriptives

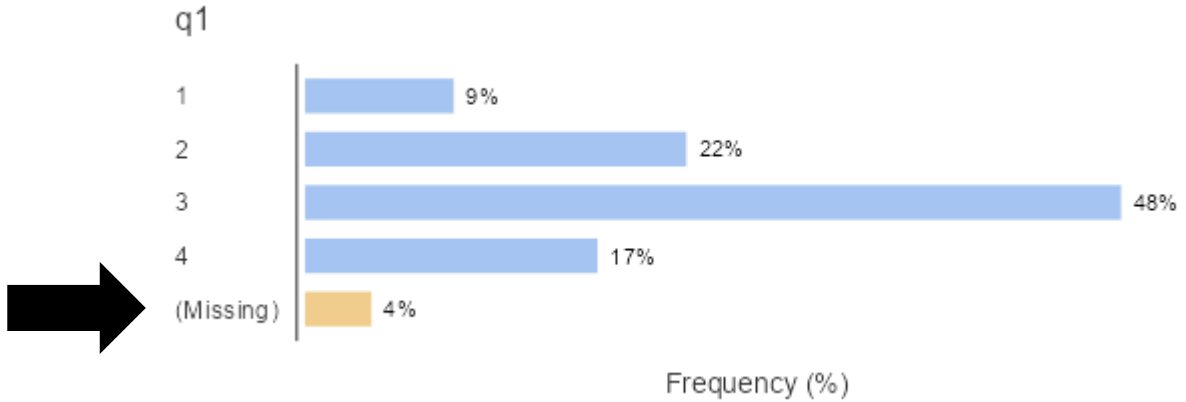
Descriptives

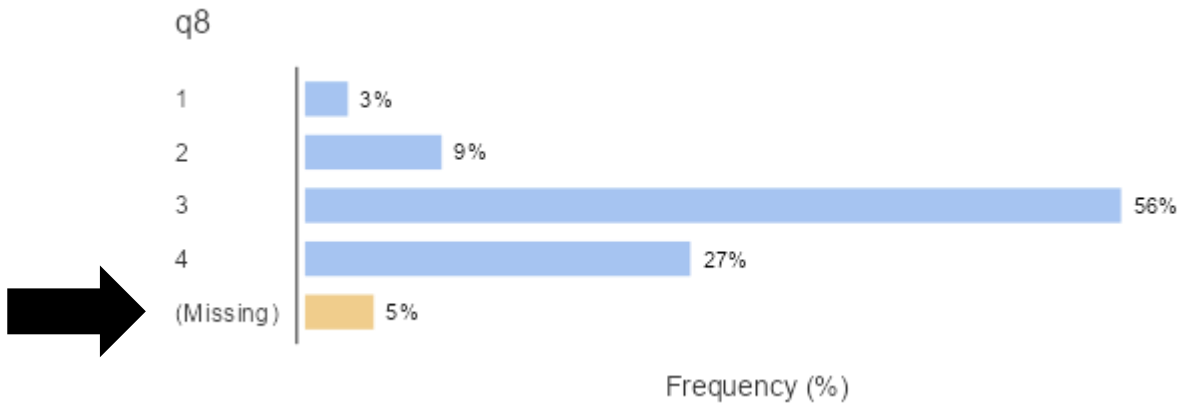
	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10
N	2050	2042	2042	2044	2043	2041	2040	2032	2046	2036
Missing	83	91	91	89	90	92	93	101	87	97
Median	3.00	3.00	3.00	3.00	3	3	3.00	3.00	2.00	2.00
Mode	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	2.00	2.00
Minimum	1	1	1	1	1	1	1	1	1	1
Maximum	4	4	4	4	4	4	4	4	4	4

Missingness



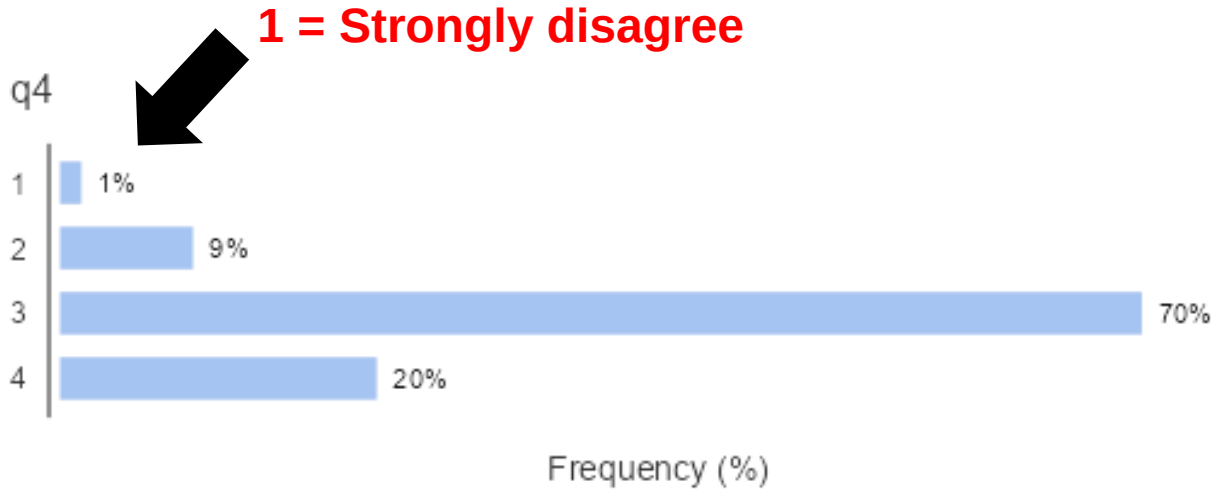
1. Install the “surveymv” module
2. Exploration → Survey Plots





Functionality

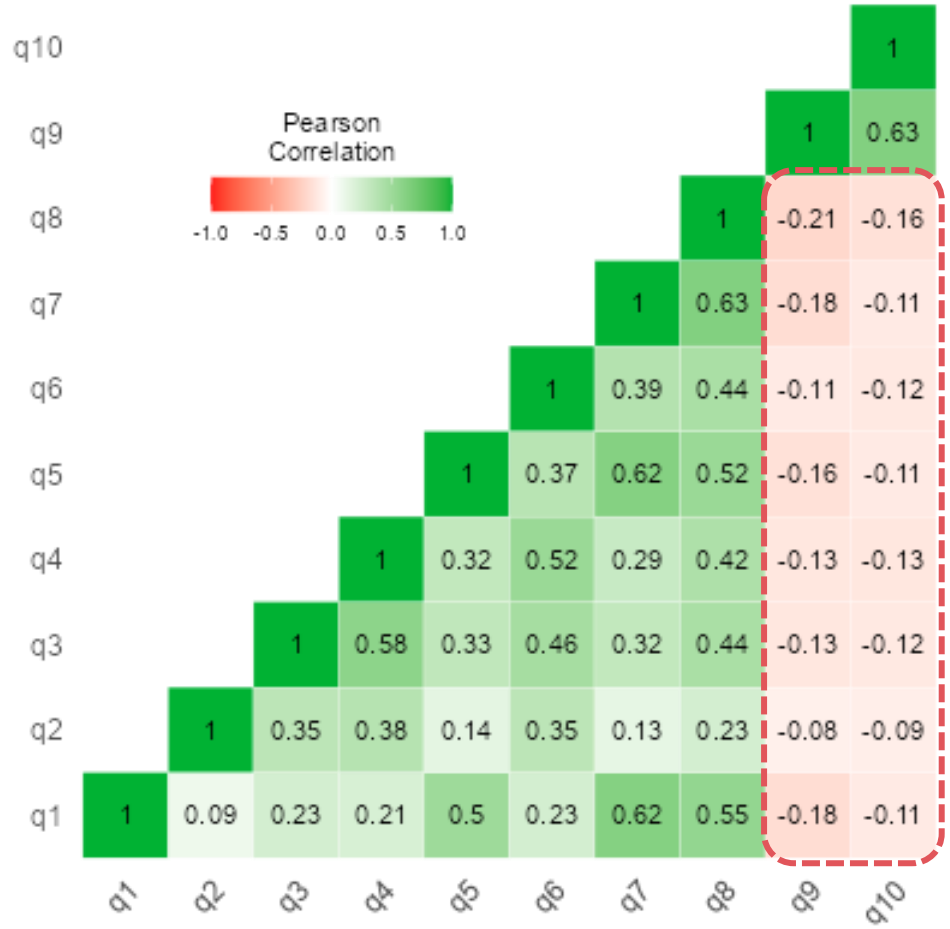
Exploration → Survey Plots



Q4. I take into account what others are interested in.

Alignment

Factor → Reliability Analysis → Correlation heatmap



Problematic items

Discrimination

☑

Factor → Reliability Analysis → Cronbach's α & Item-rest correlation

Scale Reliability Statistics

	Cronbach's α
scale	0.711

Note. items 'q9' and 'q10' correlate negatively with the total scale and probably should be reversed

[6]

It should be > 0.20

Item Reliability Statistics

	Item-rest correlation	If item dropped Cronbach's α
q1	0.4688	0.669
q2	0.2775	0.702
q3	0.4937	0.671
q4	0.4880	0.673
q5	0.5603	0.652
q6	0.5147	0.667
q7	0.5987	0.642
q8	0.6263	0.644
q9	−0.1125	0.771
q10	−0.0430	0.749

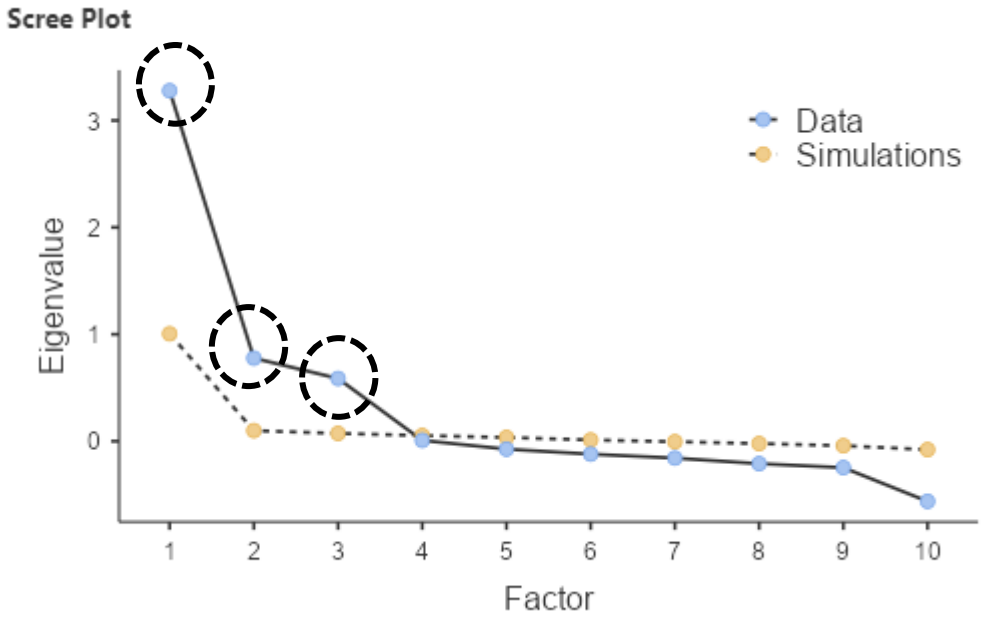
Construct Validity

Factor → Exploratory Factor Analysis → Scree plot

Factor Loadings

	Factor			Uniqueness
	1	2	3	
q1	0.774			0.463
q2		0.545		0.744
q3		0.693		0.479
q4		0.799		0.381
q5	0.659			0.490
q6		0.605		0.532
q7	0.868			0.264
q8	0.615			0.398
q9			0.794	0.357
q10			0.792	0.381

Note. 'Principal axis factoring' extraction method was used in combination with a 'oblimin' rotation



1. I prefer working as part of a team to working alone.
2. I am a good listener.
3. I enjoy seeing my classmates be successful.
4. I take into account what others are interested in.
5. I find that teams make better decisions than individuals.
6. I enjoy considering different perspectives.
7. I find that teamwork raises my own efficiency.
8. I enjoy cooperating with peers.
9. I make friends easily at school.
10. Other students seem to like me.

Construct Validity

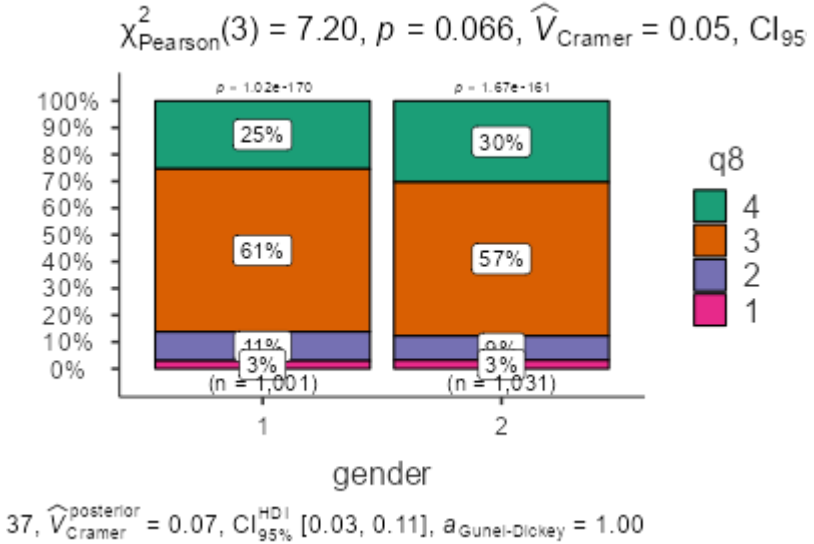
1. Frequencies → χ^2 test of association (under “Independent Samples”)
2. Install the “JJStatsPlot” module → Bar Charts

Contingency Tables

	gender		
q8	1	2	Total
1	30	33	63
2	108	93	201
3	609	592	1201
4	254	313	567
Total	1001	1031	2032

χ^2 Tests

	Value	df	p
χ^2	7.20	3	0.066
N	2032		



Null hypothesis:

There is no relationship between “Q8. I enjoy cooperating with peers” and gender.

Thank You

For questions and comments: **bulut@ualberta.ca**