Affaires mondiales Canada | Global Affairs Canada

Canada

CANADIAN FOREIGN SERVICE INSTITUTE | L'INSTITUT CANADIEN DU SERVICE EXTÉRIEUR

Introduction to Data Analysis

DATA VISUALIZATION BASICS

Patrick Boily

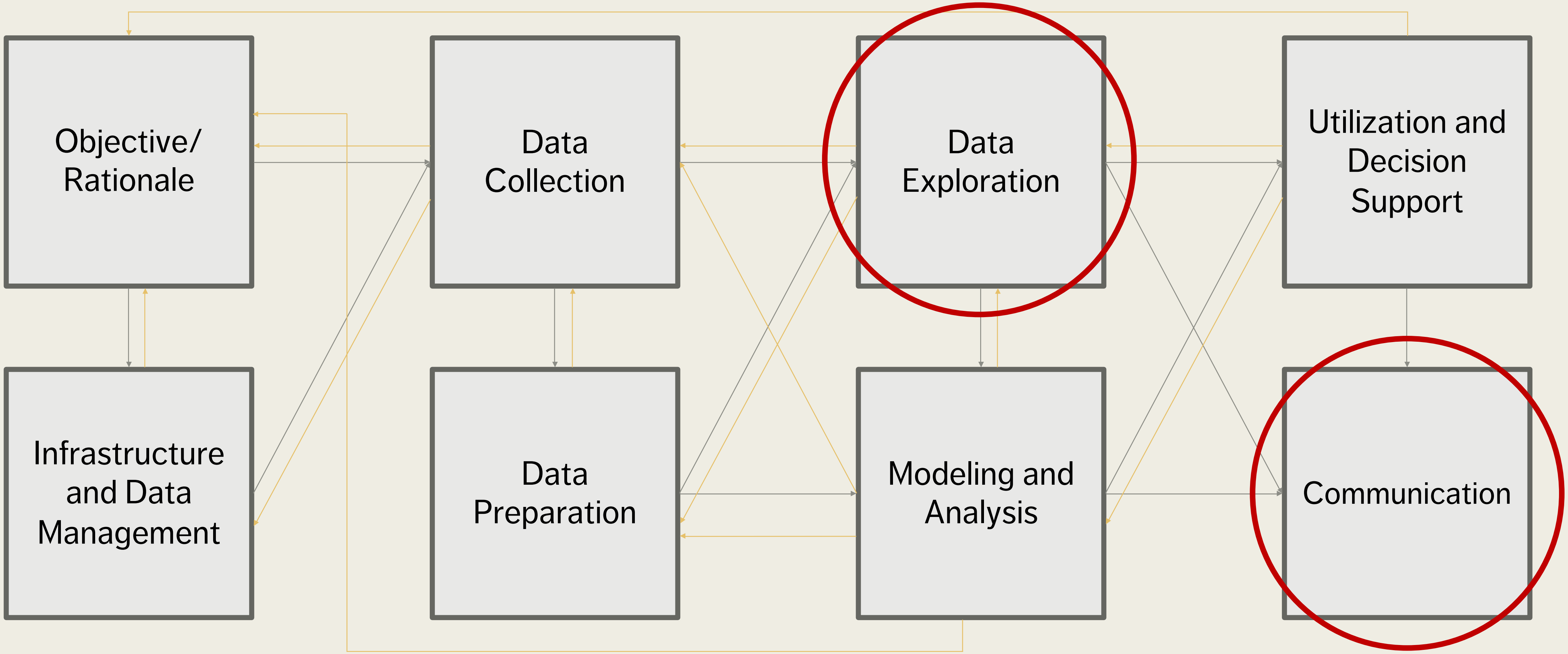
Data Action Lab | uOttawa | Idlewyld Analytics

pboily@uottawa.ca

[with files from Stephen Davies | DAVHILL]

> “Discovery is no longer limited by the collection and processing of data, but rather management, analysis, and visualization.”
>
> @DamianMingle

THE (MESSY) ANALYSIS PROCESS



Objective/ Rationale

Data Collection

Data Exploration

Utilization and Decision Support

Infrastructure and Data Management

Data Preparation

Modeling and Analysis

Communication

PRE-ANALYSIS DATA VISUALIZATION

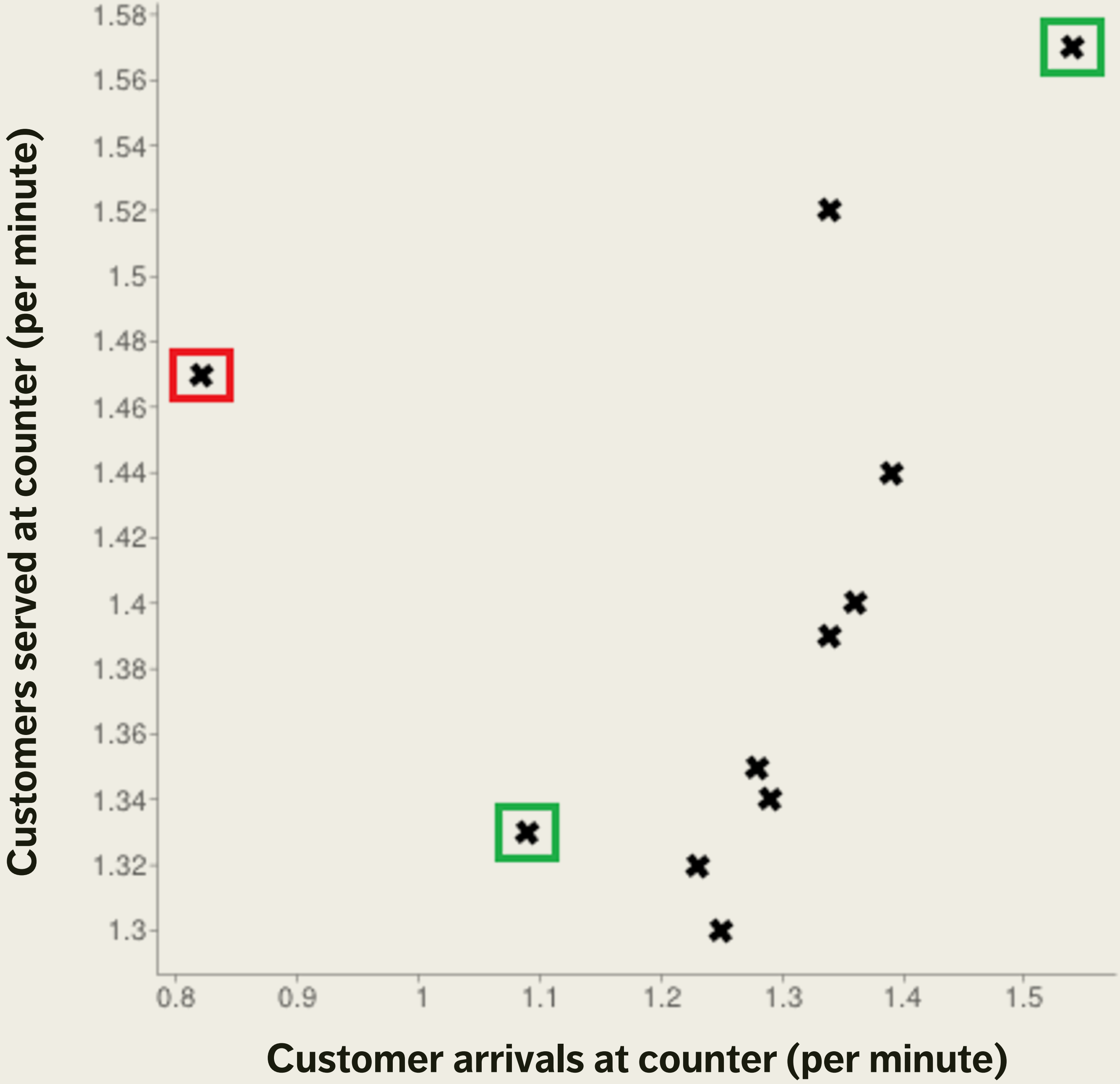
DATA VISUALIZATION BASICS

PRE-ANALYSIS USE

Data visualization can be used to set the stage for analysis:

- **detecting anomalous entries**
  invalid entries, missing values, outliers
- **shaping the data transformations**
  binning, standardization, Box-Cox transformations, PCA-like transformations
- **getting a sense for the data**
  data analysis as an art form, exploratory analysis
- **identifying hidden data structure**
  clustering, associations, patterns informing the next stage of analysis

[Personal dataset]



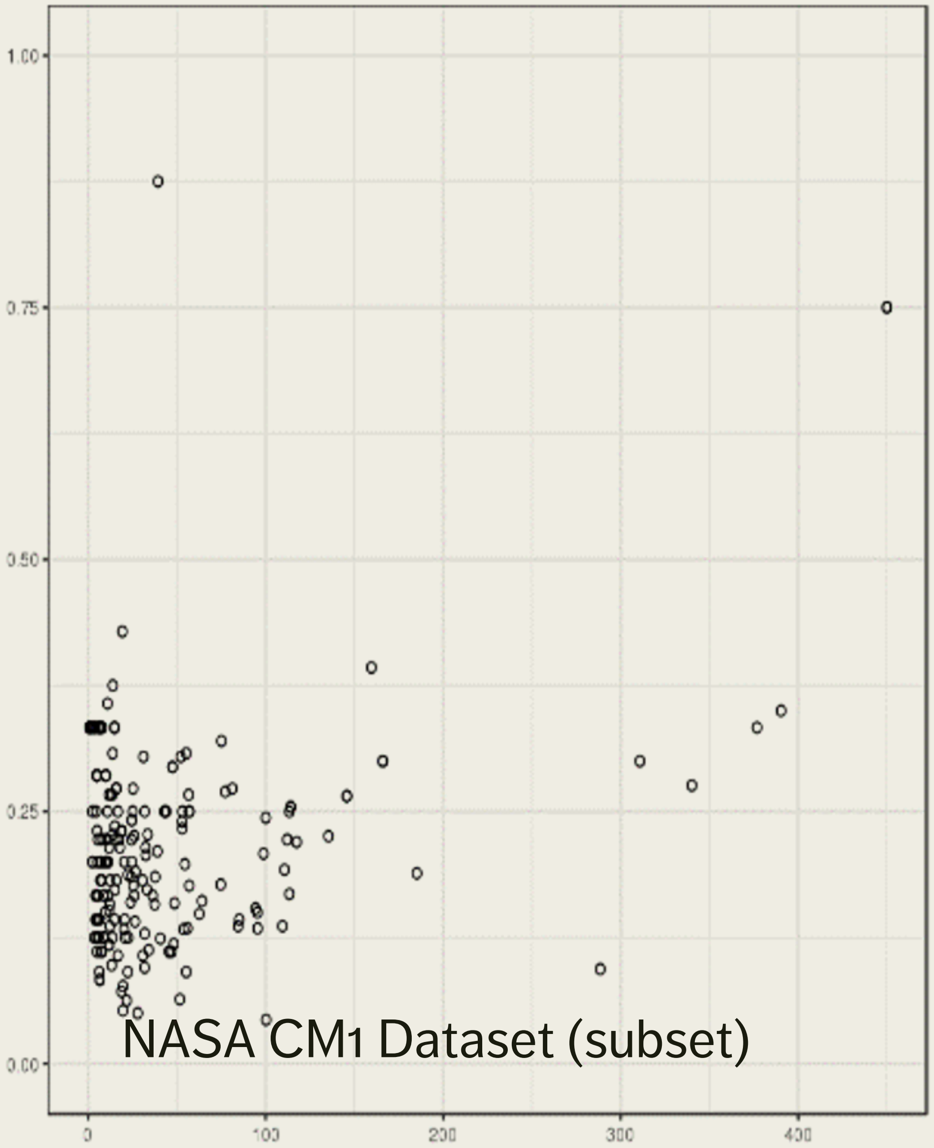
[http://promise.site.uottawa.ca/SERepository/datasets/cm1.arff]

REPRESENTING OBSERVATIONS

2 variables can be represented by position in the plane.

Additional factors can be depicted through:

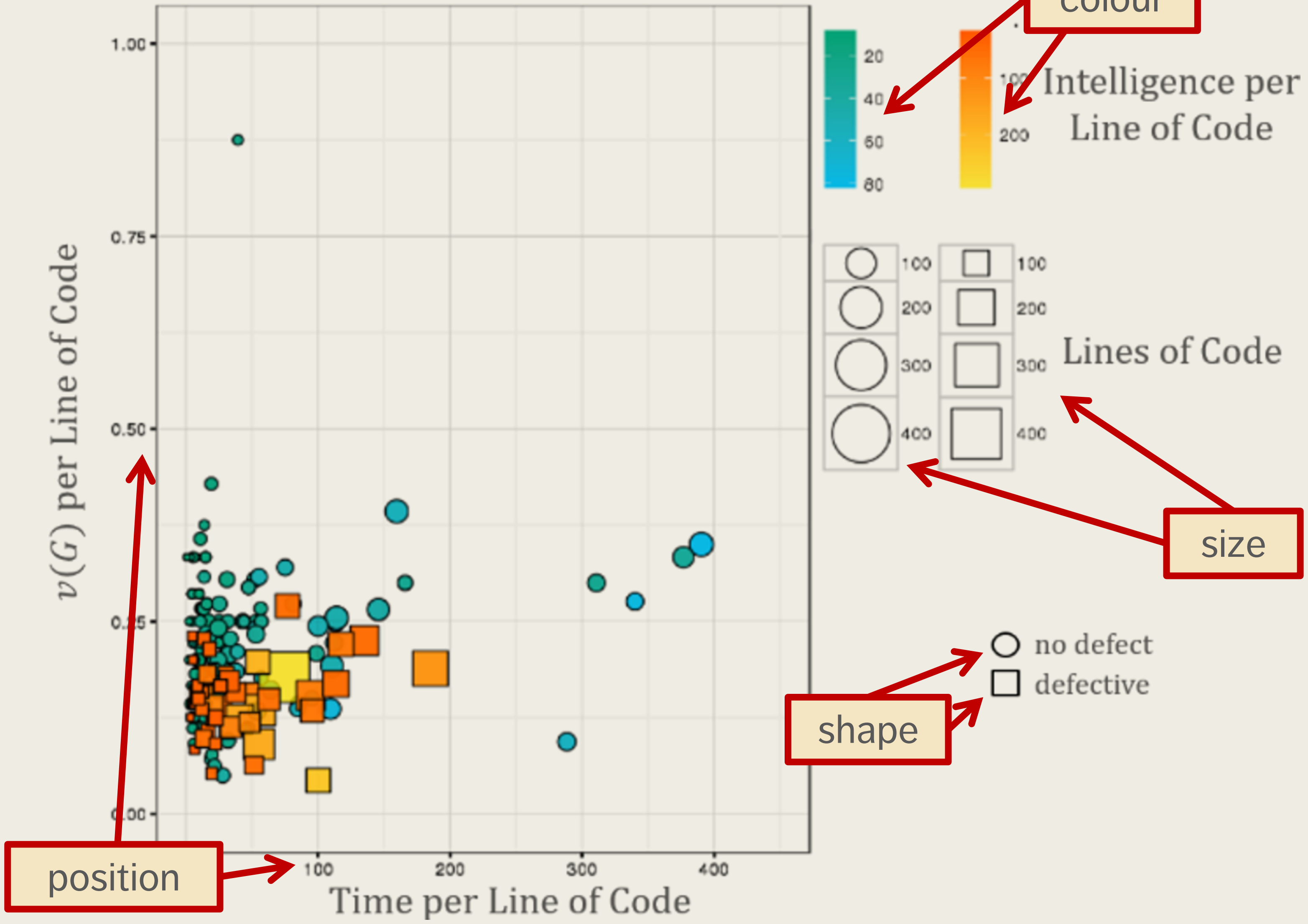
- size
- color
- value
- texture
- line orientation
- shape
- (motion?)



NASA CM1 Dataset (subset)

[http://promise.site.uottawa.ca/SERepository/datasets/cm1.arff]

NASA CM1 Dataset (subset)



colour

size

shape

position

WORKHORSE VISUALIZATIONS

Line Chart/Rug Chart/Number Line

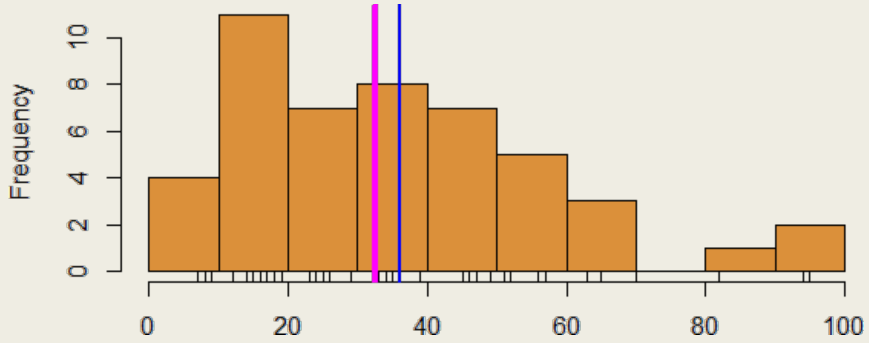
Histogram

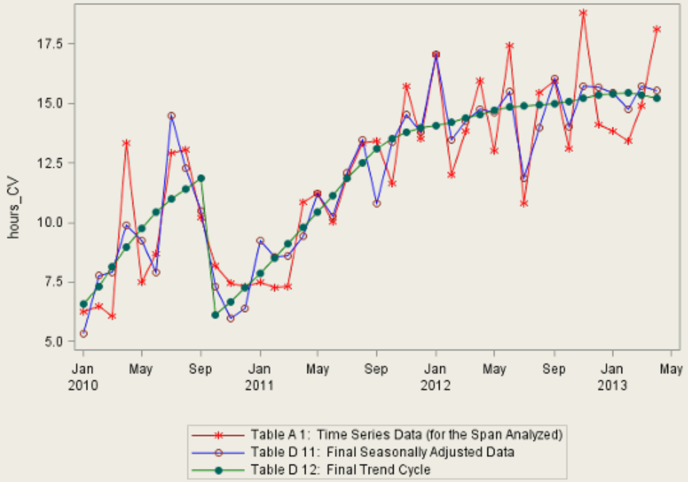
Line Graph

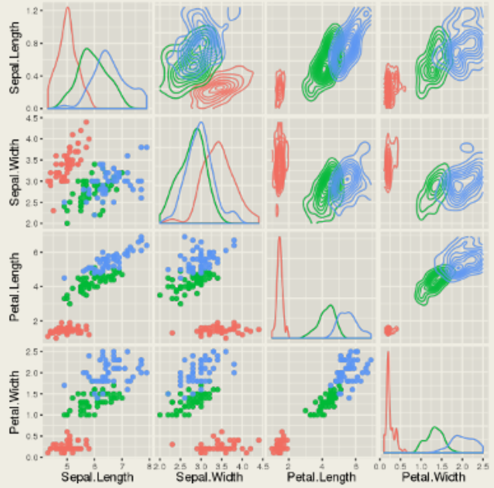
Boxplot

Bar Chart

Scatterplot







POST-ANALYSIS DATA VISUALIZATION

DATA VISUALIZATION BASICS

[E. Tufte, *Beautiful Evidence*]

FUNDAMENTAL PRINCIPLES OF ANALYTICAL DESIGN

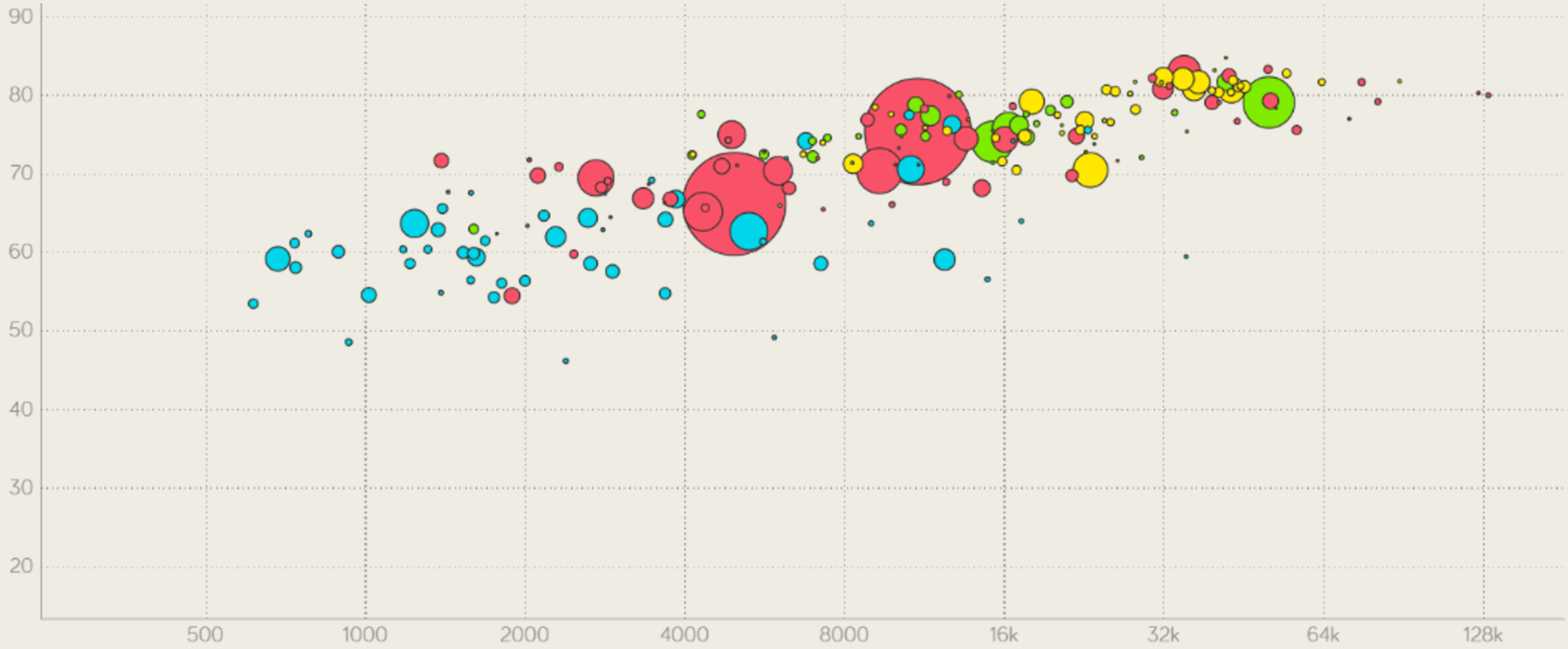
Reasoning and communicating our thoughts are intertwined with our lives in a causal and dynamic multivariate Universe.

Symmetry to visual displays of evidence: consumers should be seeking exactly what producers should be providing, namely

- meaningful comparisons
- causal networks and underlying structure
- multivariate links
- integrated and relevant data
- honest documentation
- primary focus on content

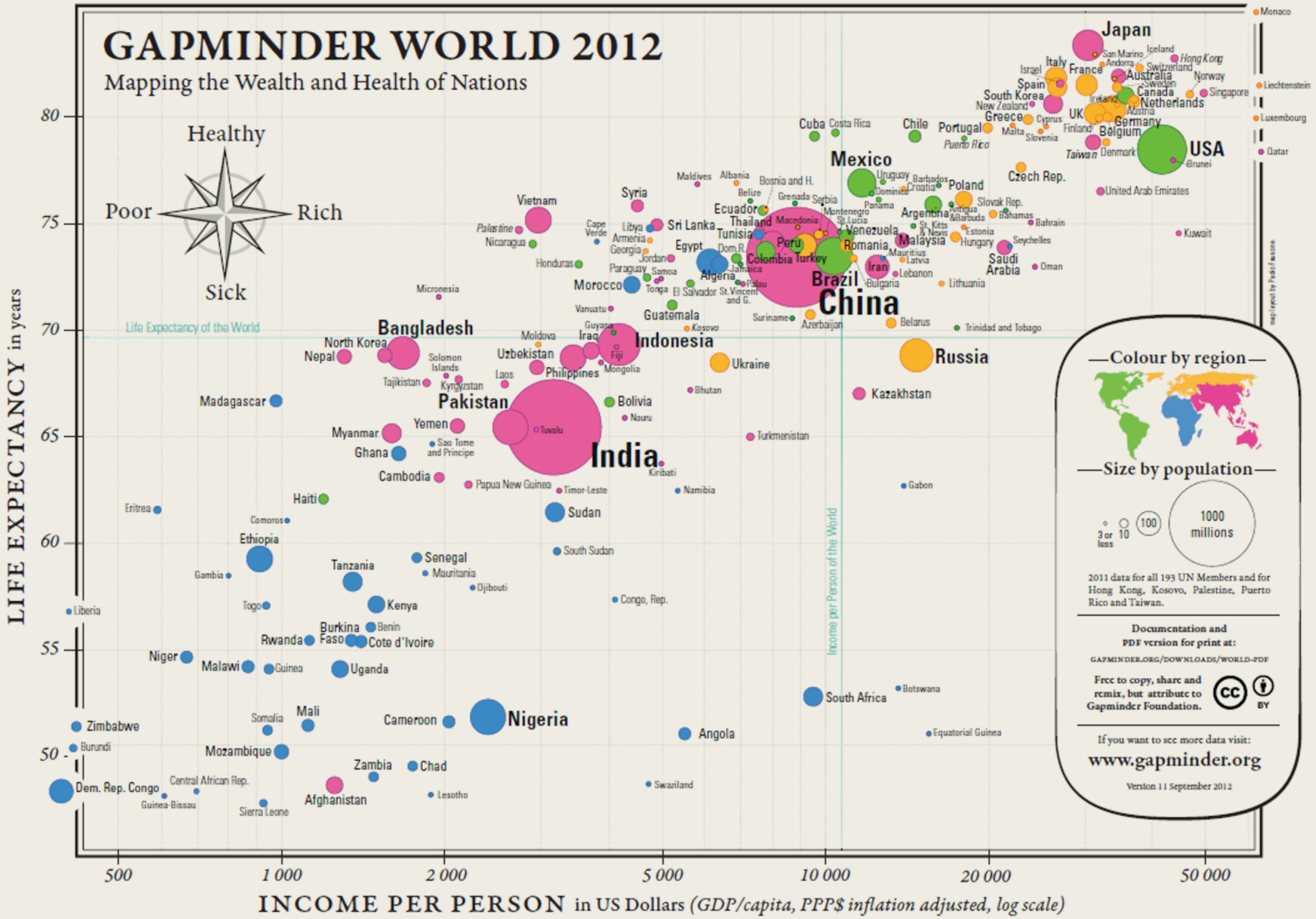
Is the point getting across?
Is the message being conveyed?

[Gapminder Foundation]



Non-Integrated Data

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



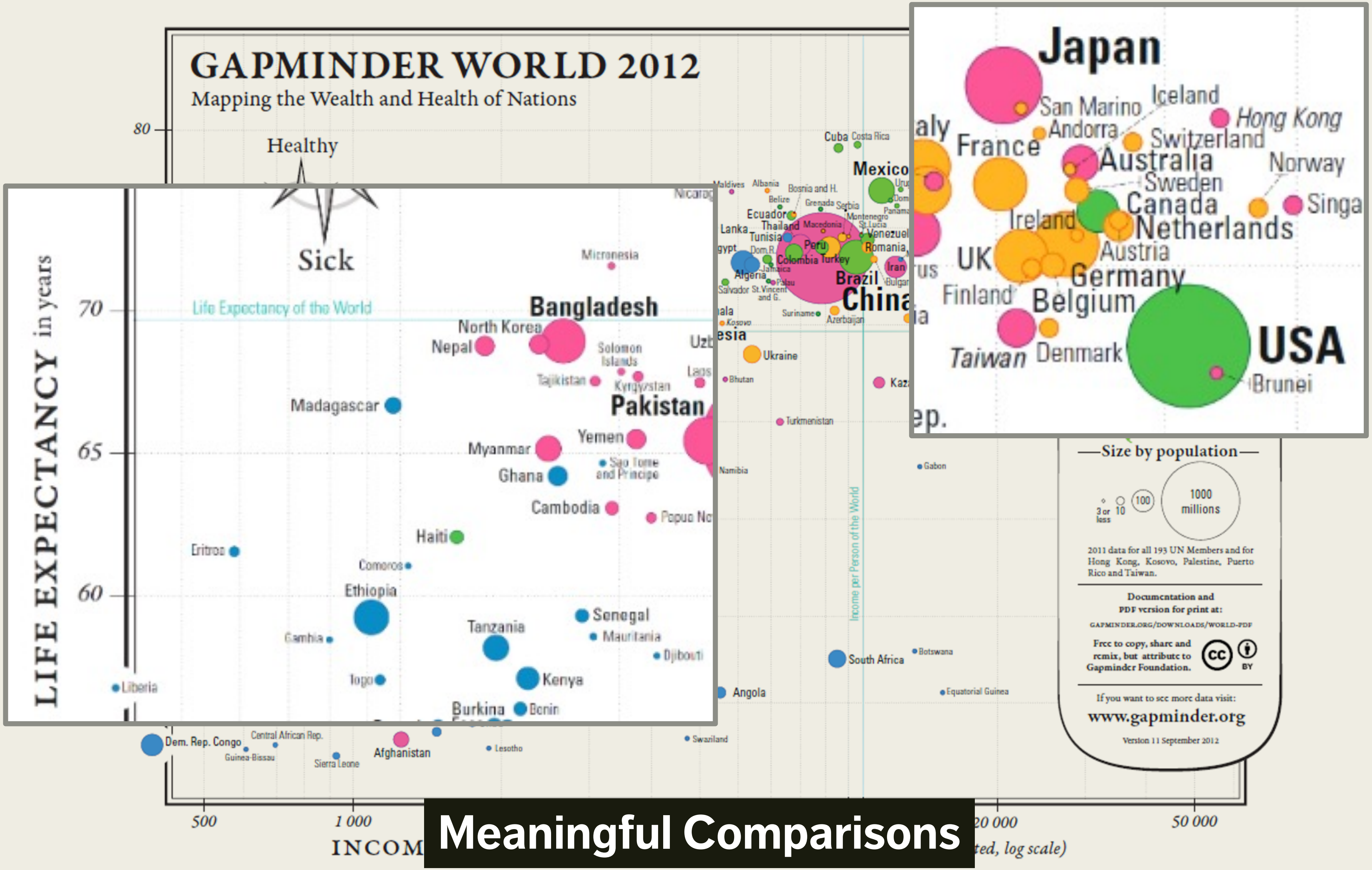
2011 data for all 193 UN Members and for Hong Kong, Kosovo, Palestine, Puerto Rico and Taiwan.

Documentation and PDF version for print at:
GAPMINDER.ORG/DOWNLOADS/WORLD-PDF

If you want to see more data visit:
www.gapminder.org

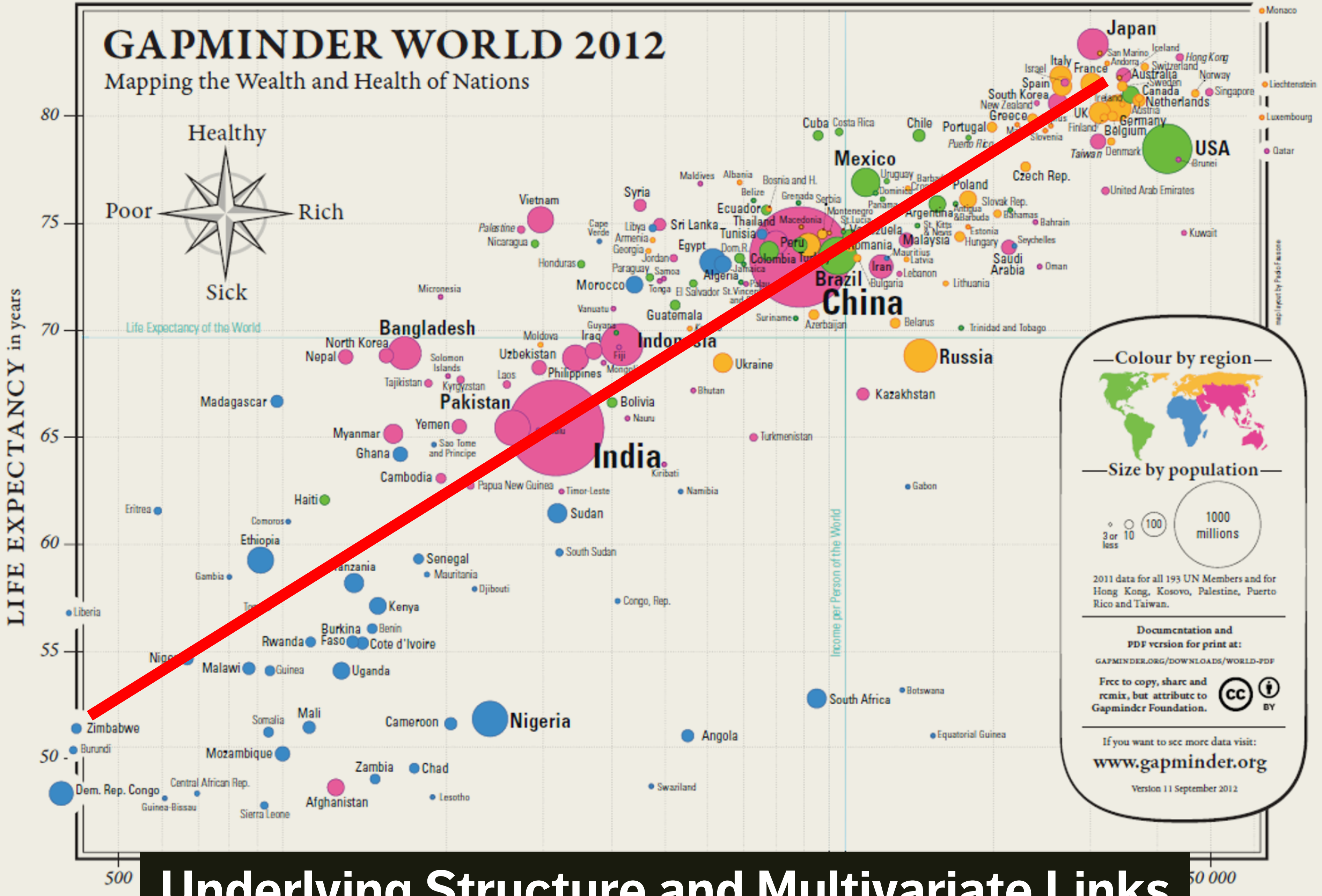
Version 11 September 2012

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



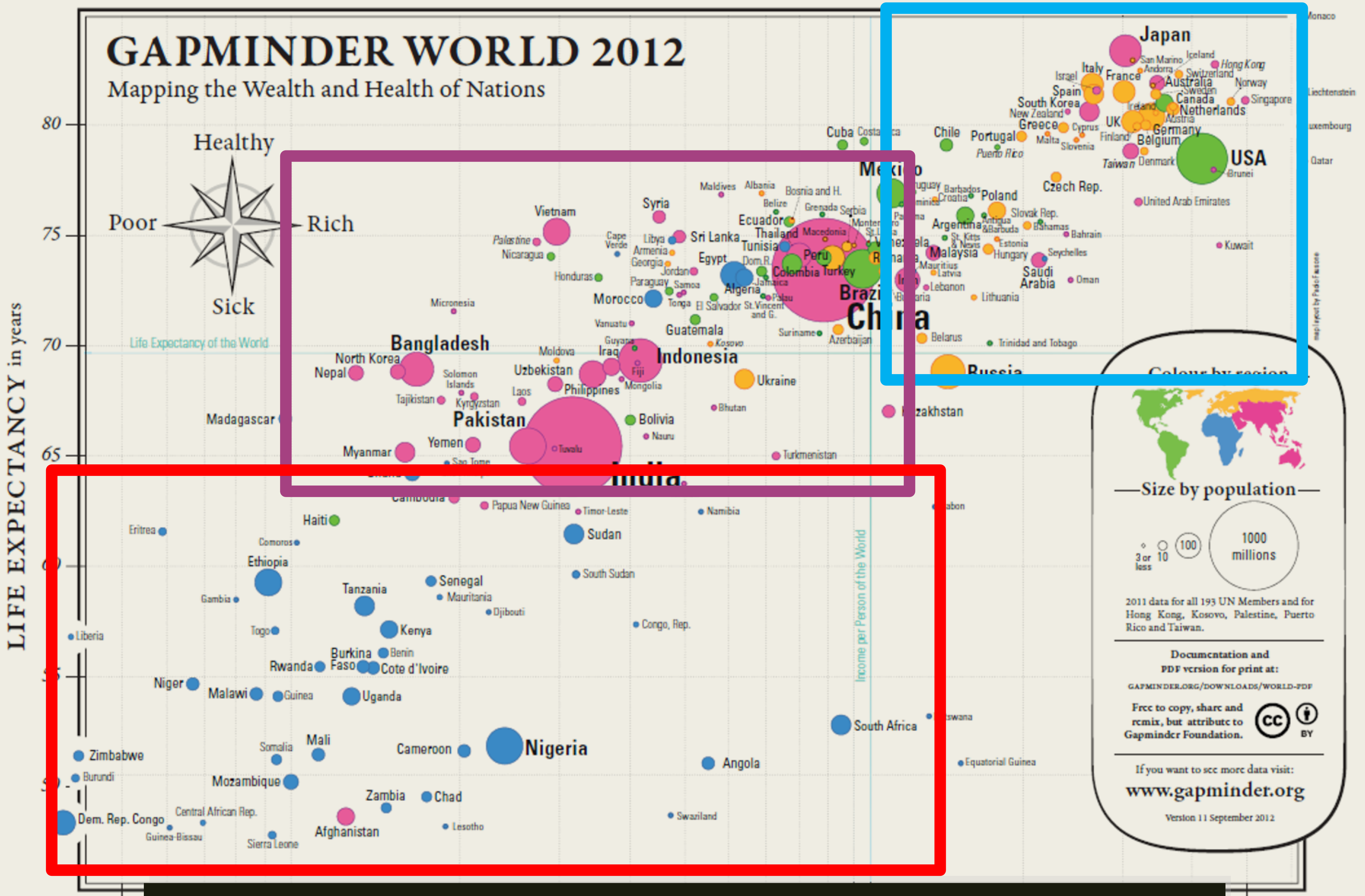
Meaningful Comparisons

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



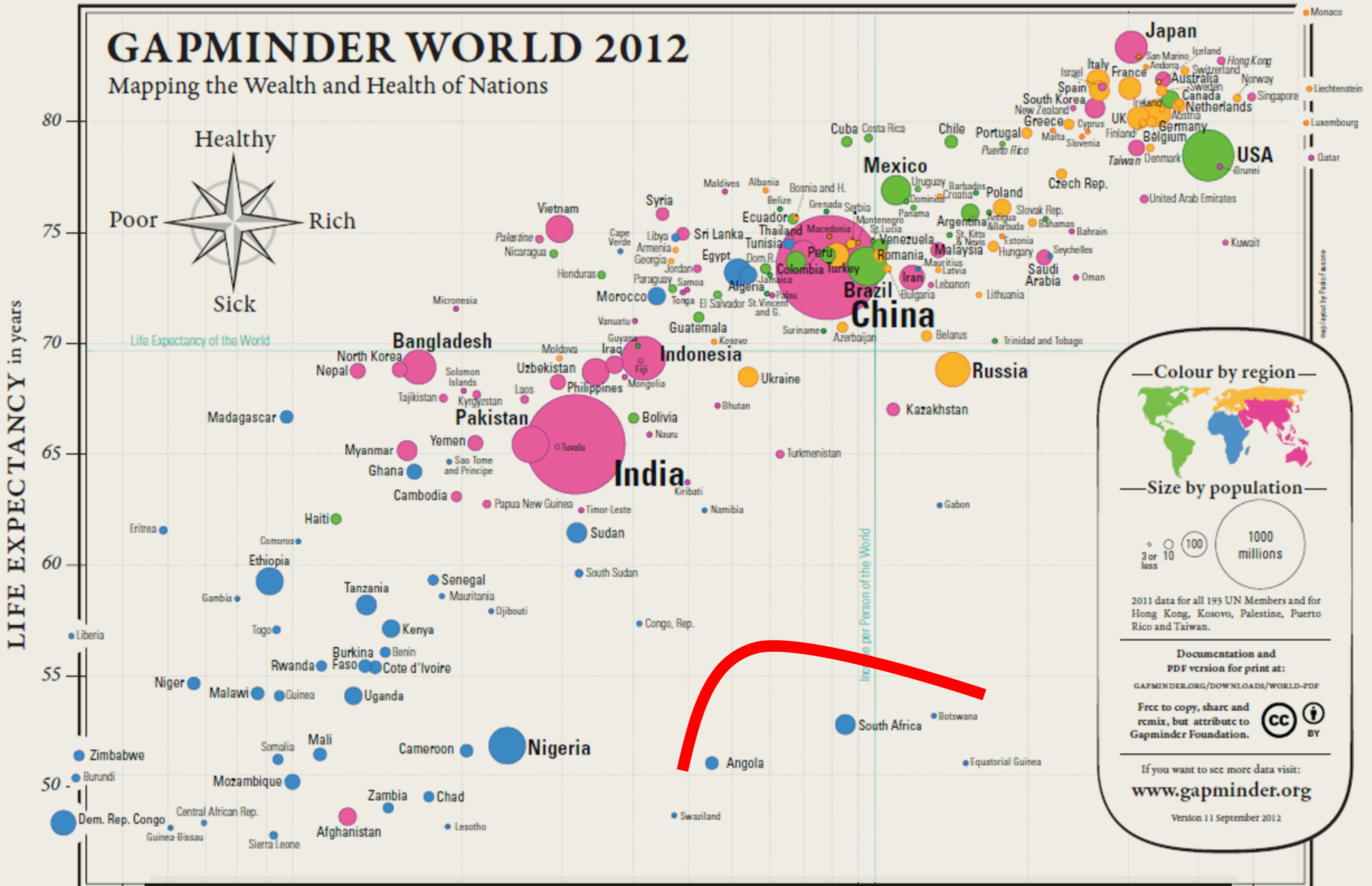
Underlying Structure and Multivariate Links

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



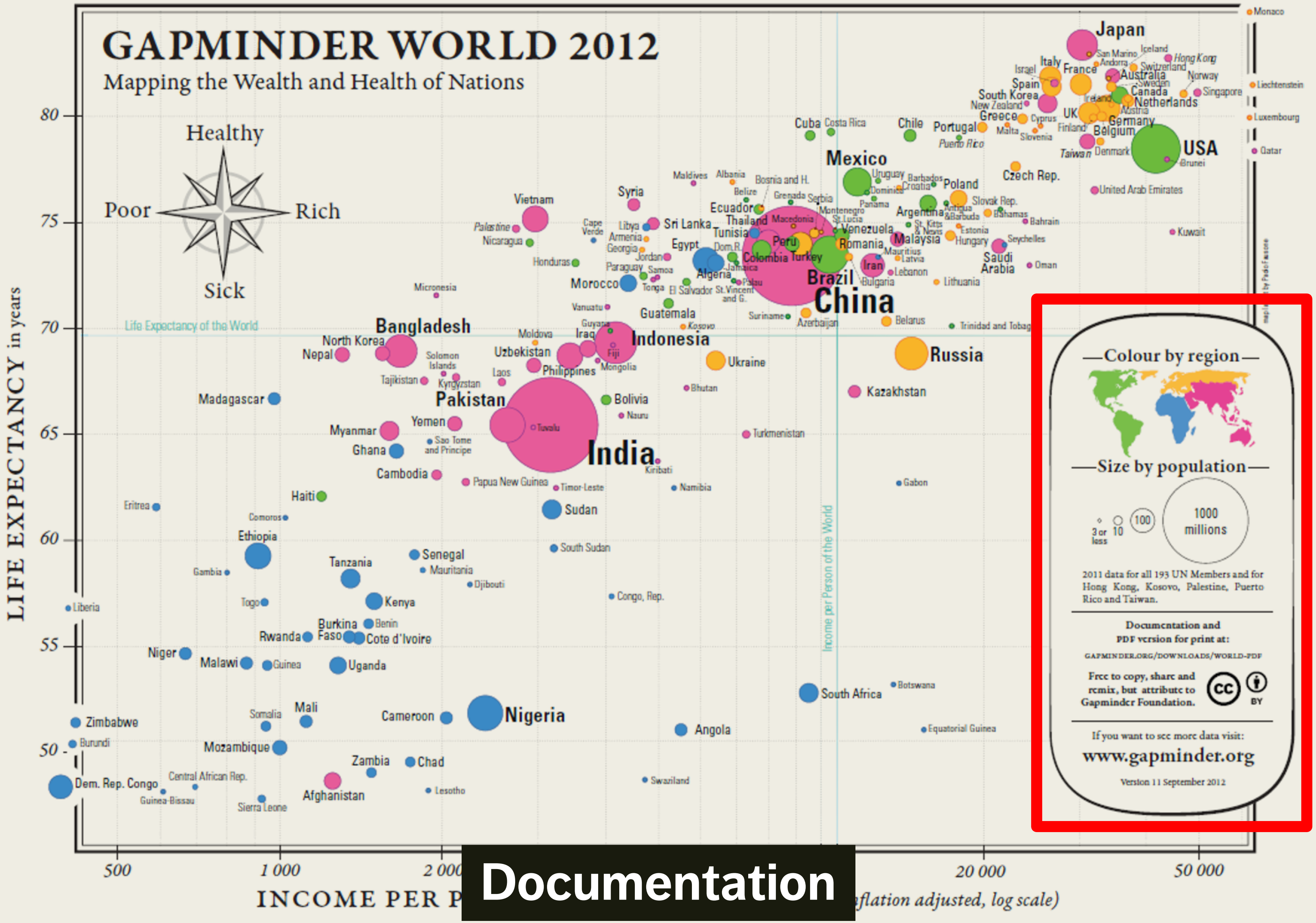
Underlying Structure and Multivariate Links

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



Underlying Structure and Multivariate Links

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



—Colour by region—

—Size by population—

3 or less 10 100 1000 millions

2011 data for all 193 UN Members and for Hong Kong, Kosovo, Palestine, Puerto Rico and Taiwan.

Documentation and PDF version for print at:
GAPMINDER.ORG/DOWNLOADS/WORLD-PDF

Free to copy, share and remix, but attribute to Gapminder Foundation. CC BY

If you want to see more data visit:
www.gapminder.org

Version 11 September 2012

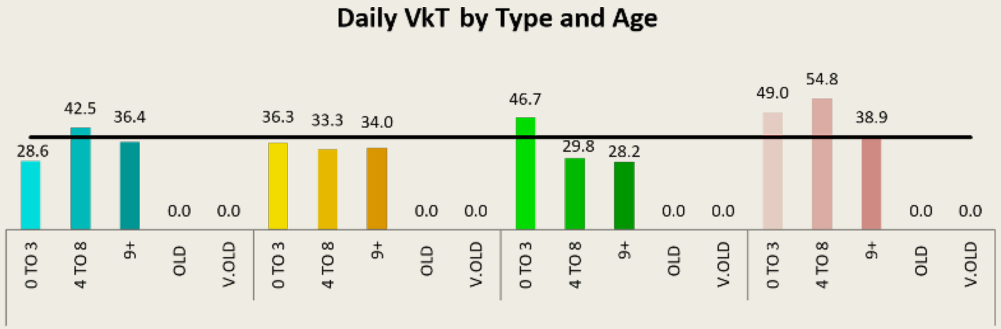
Documentation

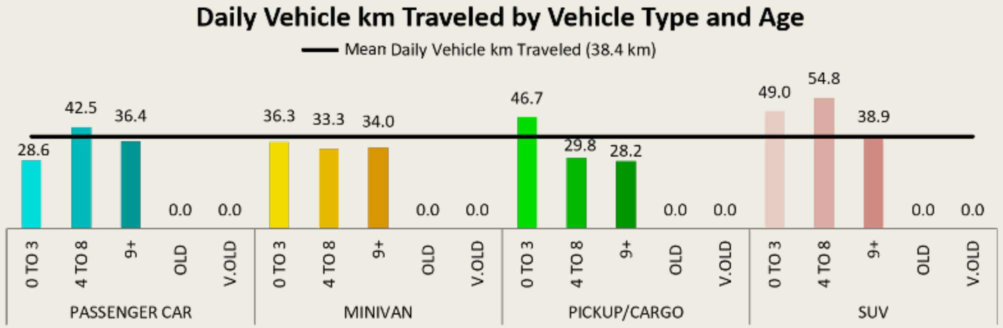
[http://flowingdata.com/tag/pitfalls]

BASIC RULES

1. Check the data
outliers, spikes, anomalies

2. Explain encoding
don't assume the reader knows what everything means

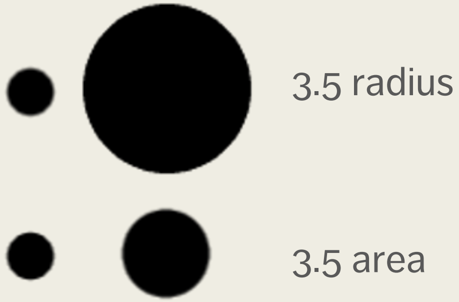




3. Label axes
knowing the scale is important

[http://flowingdata.com/tag/pitfalls]

BASIC RULES

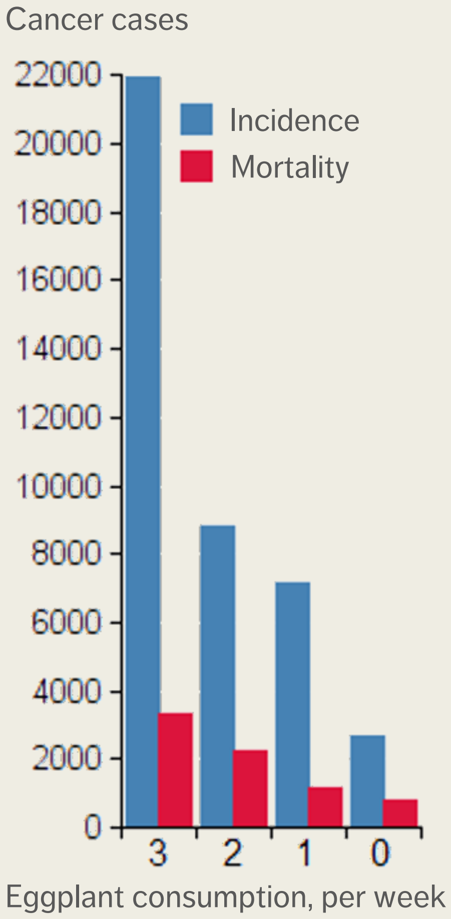


4. Include units
eliminate the need for guesswork

5. Keep your geometry in check
circles and 2D shape are sized by area, bar charts by length

6. Include your sources
protect yourself, and let those who want to dig deeper do so

7. Consider your audience
a poster can be wordy, a presentation should be minimalist



[http://dataphys.org/list/]

A WORD ABOUT ACCESSIBILITY

Charts cannot usually be translated to Braille. Describing the features and emerging structures in a visualization is a possible solution… **if they can be spotted.**

Analysts must produce clear and meaningful visualizations, but they must also describe them and their features in a fashion that allows all to "see" the insights. This requires analysts to have “seen” all the insights, which is not always possible.

Conditions: colourblindness, low vision, motor impairment, cognitive disability, ADHD, etc.

Best Practices: high contrast text/elements, zoom/magnifications, keyboard navigation, assistive design, short summaries, undo/redo functionality, etc. [F. Elavsky]

A WORD ABOUT ACCESSIBILITY

Data Perception:

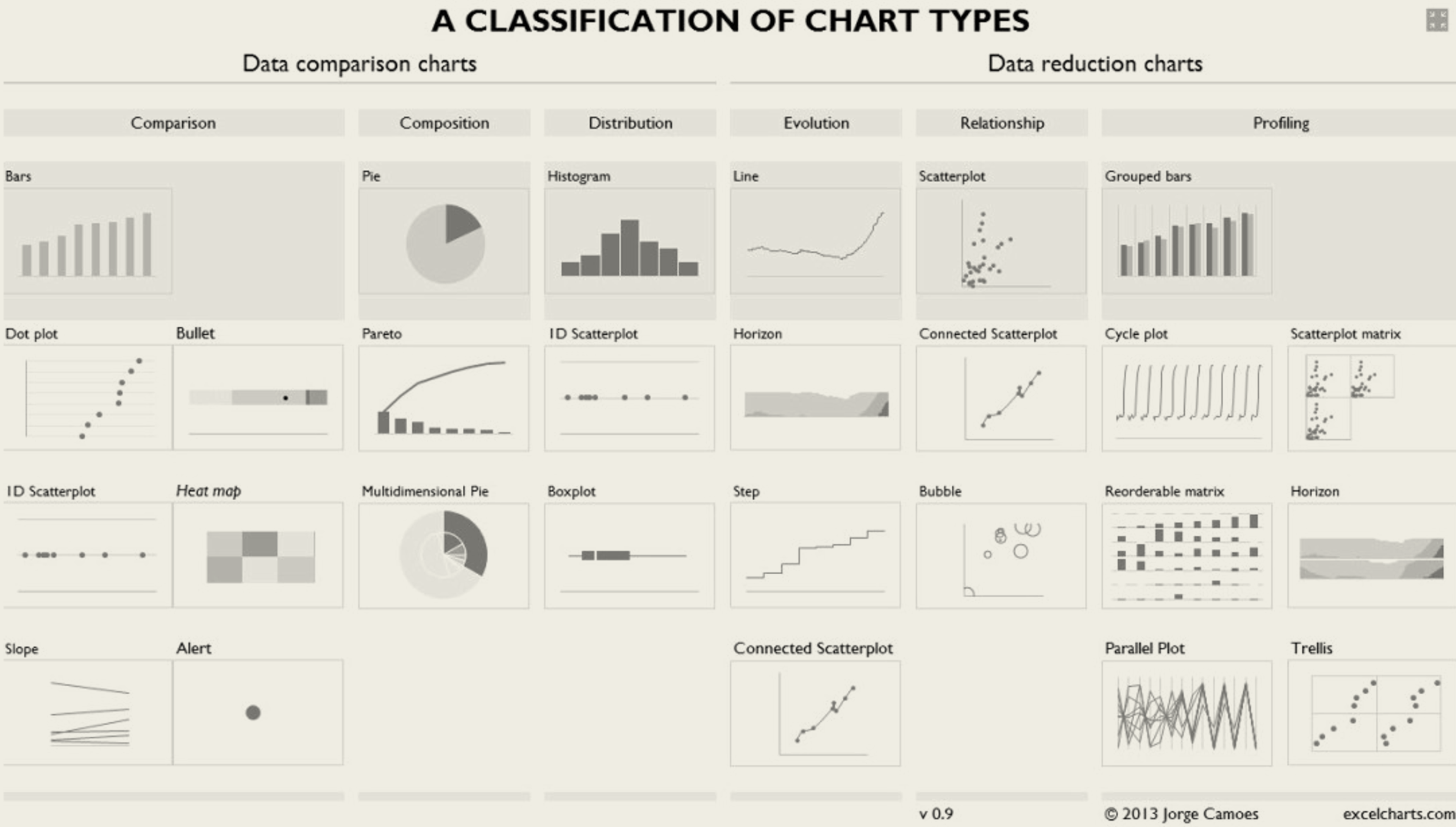
- texture-based representations
- text-to-speech
- sound/music
- odor-based or taste-based representations (?!?)

Sonifications:

- TRAPPIST Sounds : TRAPPIST-1 Planetary System Translated Directly Into Music
- Listening to data from the Large Hadron Collider, L. Asquith

VISUALIZATION CATALOGUE

DATA VISUALIZATION BASICS



A CLASSIFICATION OF CHART TYPES

Data comparison charts				Data reduction charts			
Comparison		Composition	Distribution	Evolution	Relationship	Profiling	
Bars		Pie	Histogram	Line	Scatterplot	Grouped bars	
Dot plot	Bullet	Pareto	1D Scatterplot	Horizon	Connected Scatterplot	Cycle plot	Scatterplot matrix
1D Scatterplot	*Heat map*	Multidimensional Pie	Boxplot	Step	Bubble	Reorderable matrix	Horizon
Slope	Alert			Connected Scatterplot		Parallel Plot	Trellis

v 0.9

excelcharts.com

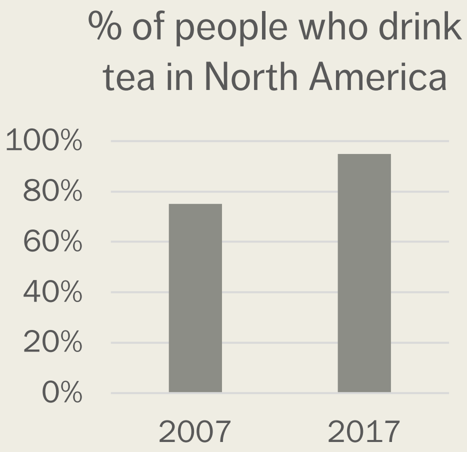
DATA DISPLAYS

With data displays, we try to highlight:

1. a **relationship** (show a connection or correlation between two or more variables);
2. a **comparison** (set some variables apart from others, and display how those two variables interact);
3. a **composition** (collect different types of information that make up a whole and display them together), and
4. a **distribution** (lay out a collection of related or unrelated information to see how it correlates, if at all, and to understand if there's any interaction between the variables).

SIMPLE TEXT AND TABLES

One or two numbers to focus on may help “set the scene” and **draw focus** to an area of the report.



95% of the population drinks tea today compared to 75% in 2007

Tables interact with our **verbal** system (we **read** them):

- used to **compare** values
- audiences will look for **their** rows

Table design needs to blend into background:

- the data should stand out, not the borders
- dense table: use **alternating** row colour

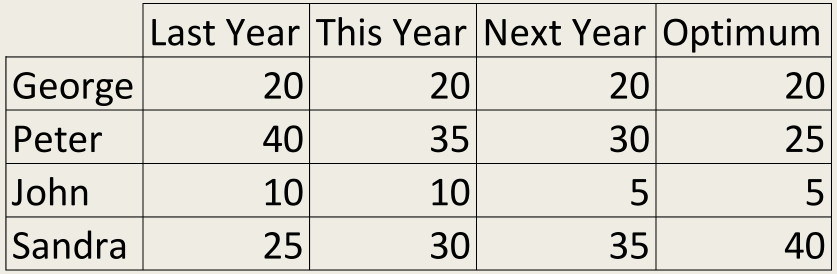
Leverage colour to convey magnitude:

- use **single colour saturation**
- use a legend to remove values

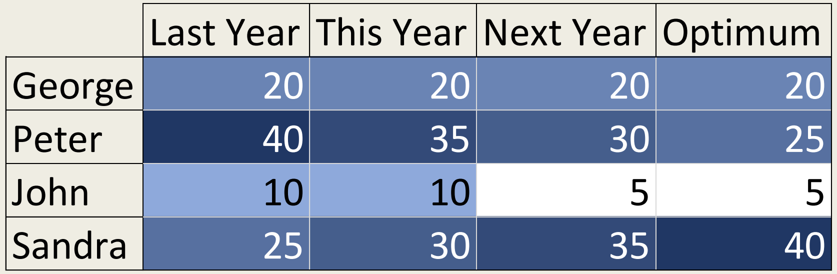
TABLES AND TABLE HEATMAPS

Name	Last Year	This Year
Ron	20	30
Fred	30	40
George	10	15

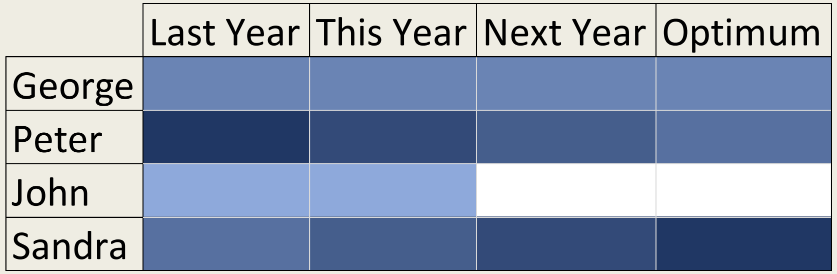
Name	Last Year	This Year
Ron	20	30
Fred	30	40
George	10	15



	Last Year	This Year	Next Year	Optimum
George	20	20	20	20
Peter	40	35	30	25
John	10	10	5	5
Sandra	25	30	35	40



	Last Year	This Year	Next Year	Optimum
George	20	20	20	20
Peter	40	35	30	25
John	10	10	5	5
Sandra	25	30	35	40



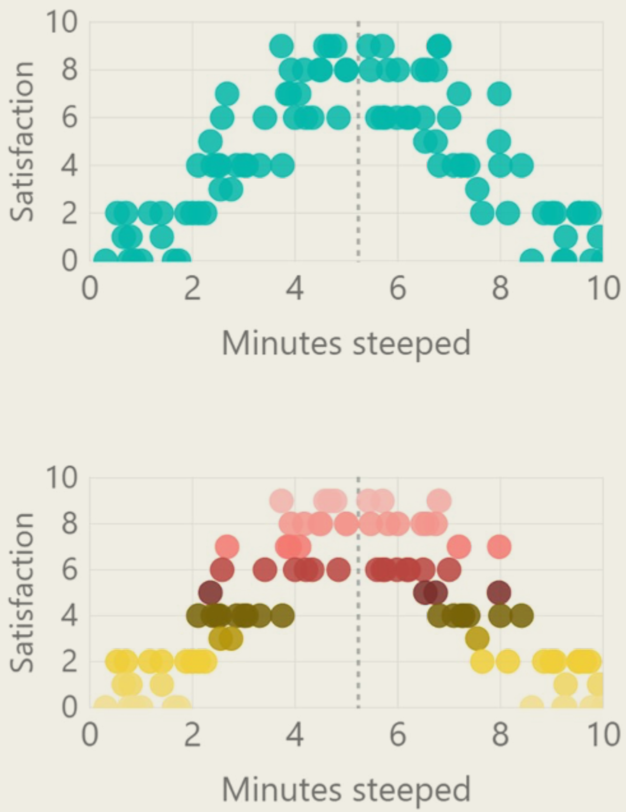
	Last Year	This Year	Next Year	Optimum
George				
Peter				
John				
Sandra				

SCATTERPLOTS

Show relationship between 2 variables (**scatterplot**) or 3 variables (**bubble plot**):

- use average lines (dotted lines) to provide context
- far fewer options in Power BI than in R or Excel
- consider using groupings to add clarity (e.g. **colour gradients**)

How long should the perfect cup of tea be steeped?



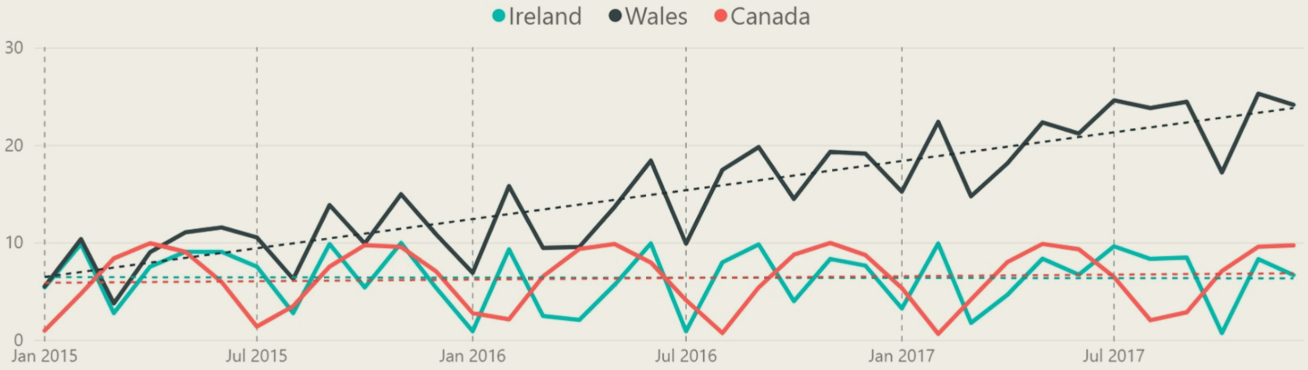
LINE CHARTS

Line chart can show a single series or multiple series of data (particularly useful for **time series**).

Axis scale should be **clear** and **relevant**.

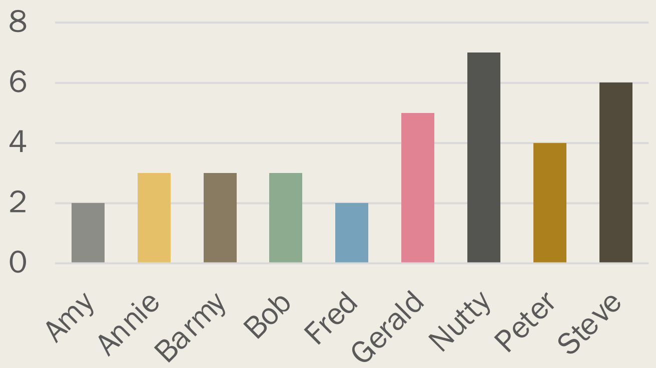
May wish to “**anchor**” y −axis if using dynamic filters

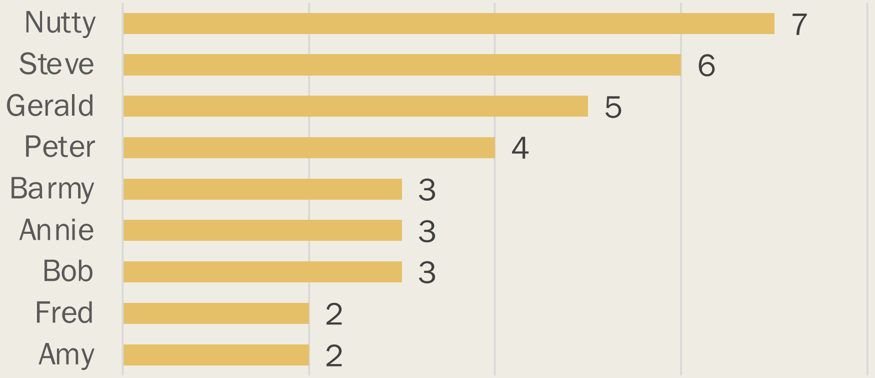
- otherwise the graph can jump around as people interact with it



Comparison of Countries – cups of tea drunk per week per person

BAR CHARTS





Versatile and useful.

ALWAYS (?) use a zero baseline.

Use graph axis OR data labels: axis for broad statements, data labels for details.

Horizontal charts are apparently **easier to read** (according to many studies).

Think about the ordering of categories.

CHART TYPES

Stacked Bar Charts

100% Bar Charts

Area Charts

Treemaps

Gauge Charts

Heatmaps and Choropleth Maps

Geographical Maps

Parallel Coordinates

Chernoff Faces

Word Clouds

Network Diagrams

Dendrograms and Trees

Sparklines

Interactive Charts

Small Multiples

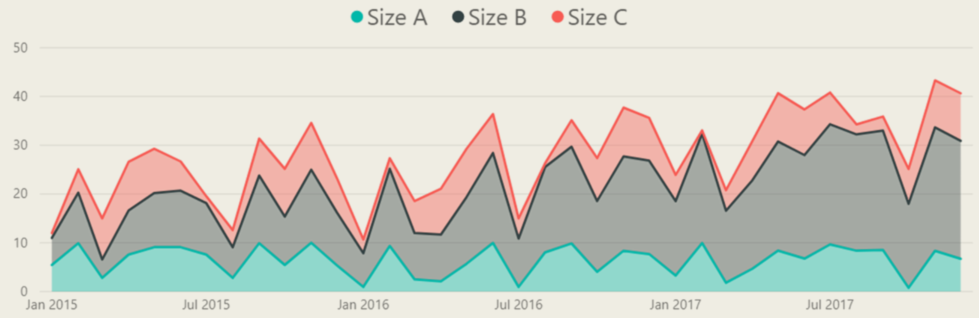
etc.

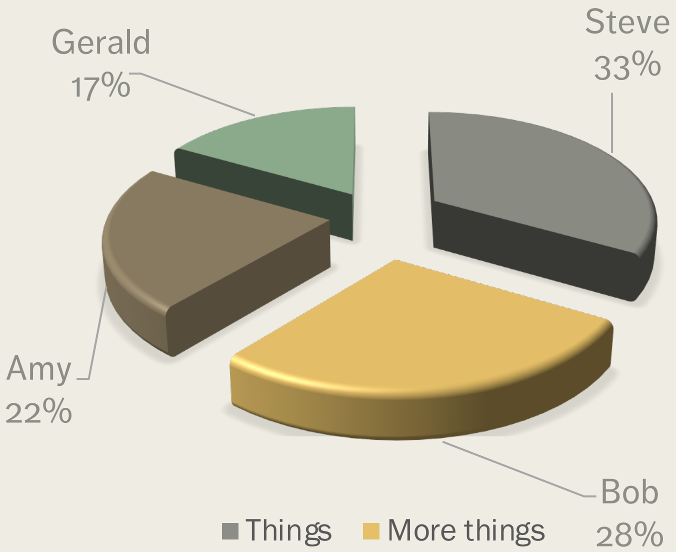
CHARTS TO AVOID

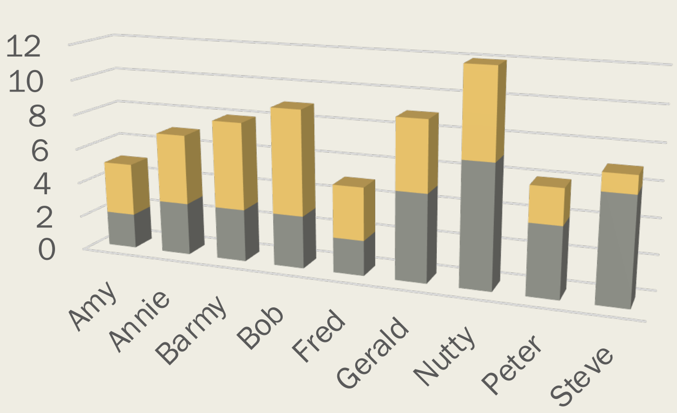
AVOID (?) anything with an arc (except gauge charts): pie, donut, etc: human brains have a hard time **comparing arcs** -- without labels, how different are Steve & Bob?

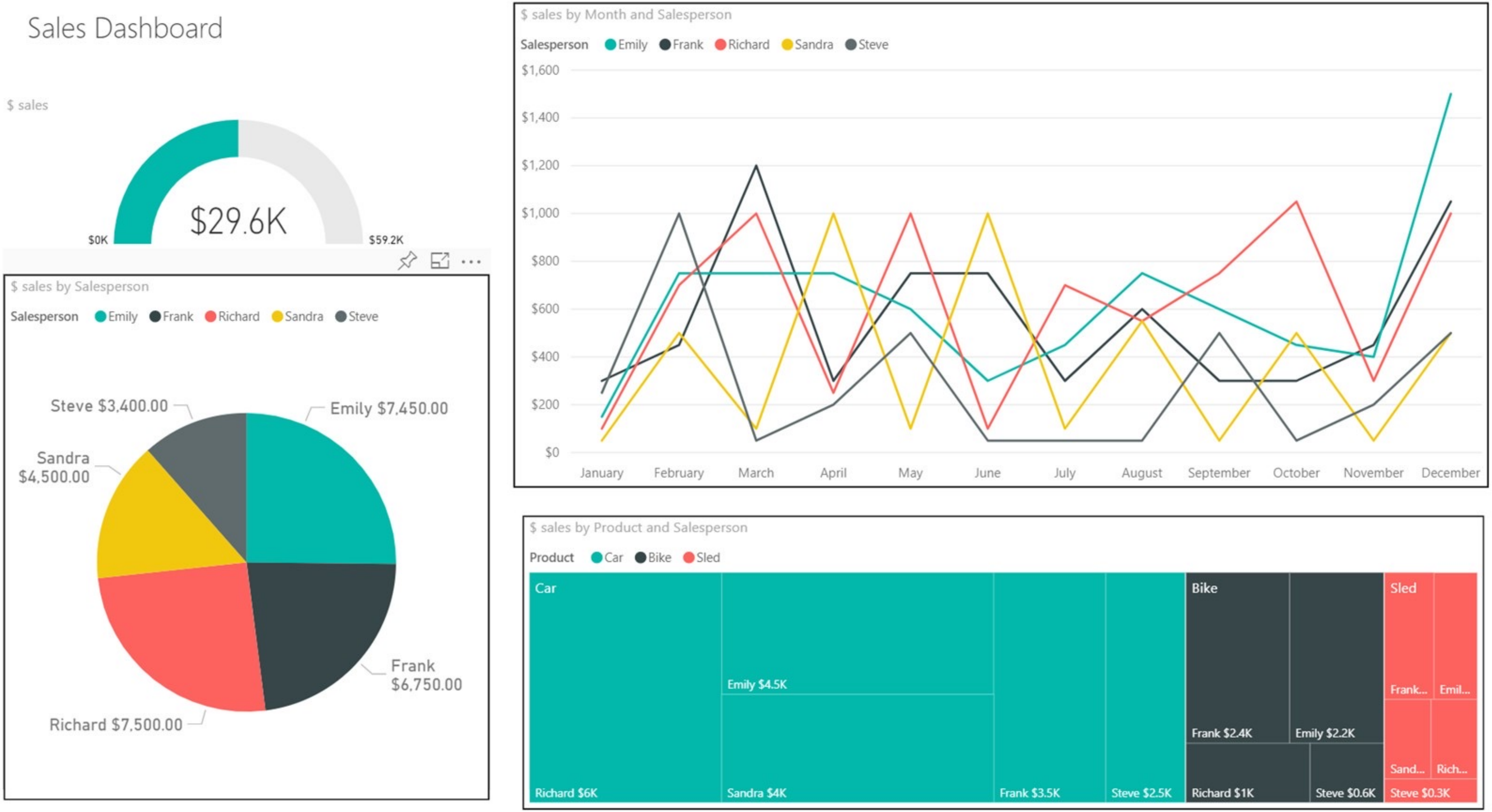
AVOID 3D charts: it is difficult to compare them visually (and they add **too much** clutter).

AVOID stacked area charts: way too confusing.









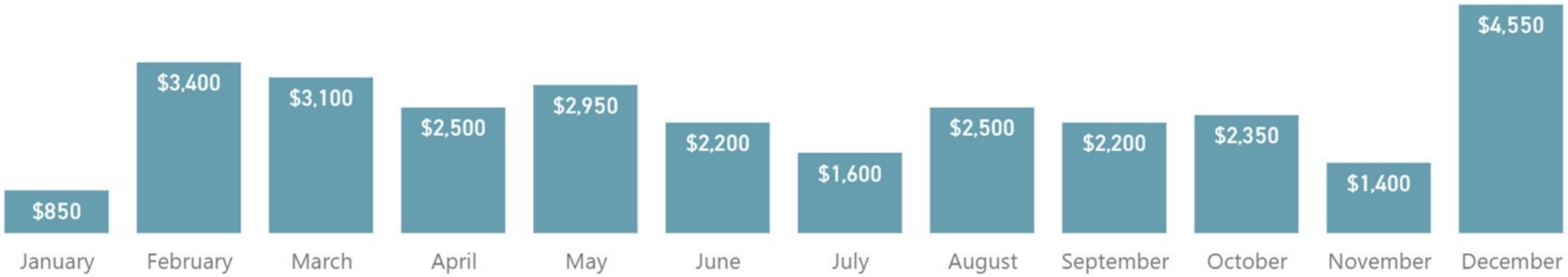
Sales Dashboard
$ sales
$29.6K
$0K
$59.2K
$ sales by Salesperson
Salesperson ● Emily ● Frank ● Richard ● Sandra ● Steve
Steve $3,400.00
Emily $7,450.00
Sandra $4,500.00
Frank $6,750.00
Richard $7,500.00
$ sales by Month and Salesperson
Salesperson ● Emily ● Frank ● Richard ● Sandra ● Steve
$1,600
$1,400
$1,200
$1,000
$800
$600
$400
$200
$0
January
February
March
April
May
June
July
August
September
October
November
December
$ sales by Product and Salesperson
Product ● Car ● Bike ● Sled
Car
Richard $6K
Emily $4.5K
Sandra $4K
Frank $3.5K
Steve $2.5K
Bike
Frank $2.4K
Emily $2.2K
Richard $1K
Steve $0.6K
Sled
Frank...
Emil...
Sand...
Rich...
Steve $0.3K

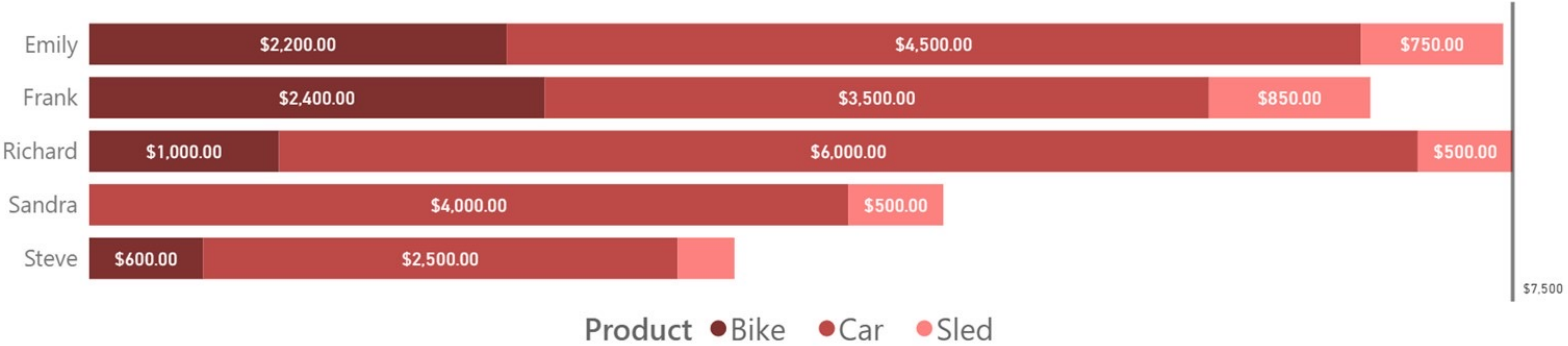
Sales Dashboard

Annual Sales for 2017

Total Sales

$29.6K





TAKE-AWAYS

Effective data visualizations **provide insights** and **facilitate understanding**.

The basic principles can guide your visualization design and consumption.

Be **creative** but keep your data and your representations **honest**.

Be mindful of attempts to distort trends and conclusions with flashy visuals.

Data and code should be made available along with the displays.

BASIC RULES OF DESIGN AND LAYOUT

DATA VISUALIZATION BASICS

VISUAL PROCESSING

Perception is **fragmented** – eyes are continuously scanning.

Visual thinking seeks patterns

- **Pre-attentive processes:** fast, instinctive, efficient, multitasking *gather information and build patterns:*

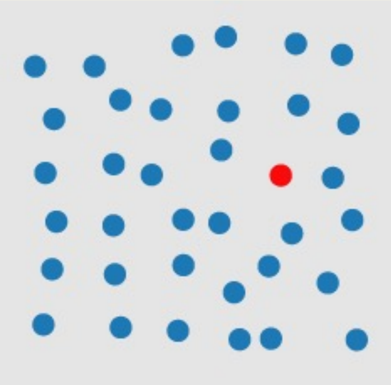
  features → patterns → objects

- **Attentive process:** slow, deliberate, focused *discover features in the patterns:*

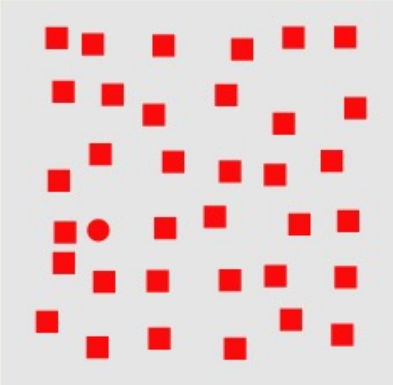
  objects → patterns → features

Challenge: highlighting one aspect of a chart can make other aspects harder to see.

pre-attentive

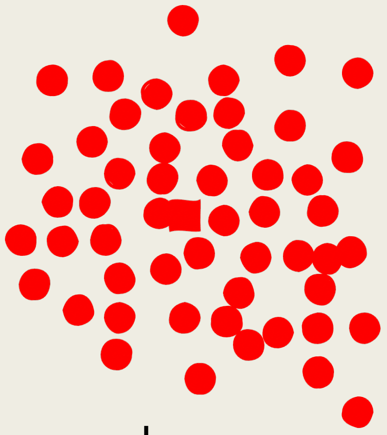


attentive

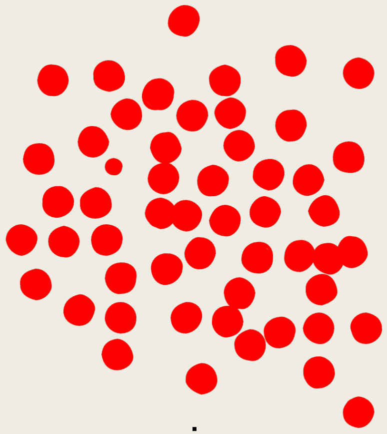


[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]

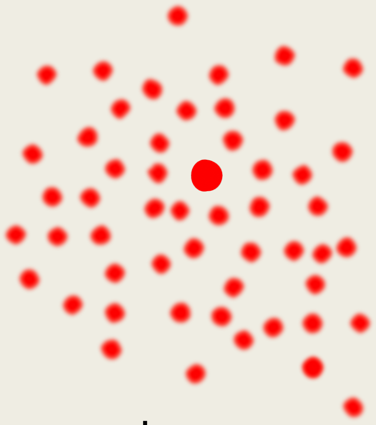
PRE-ATTENTIVE FEATURES



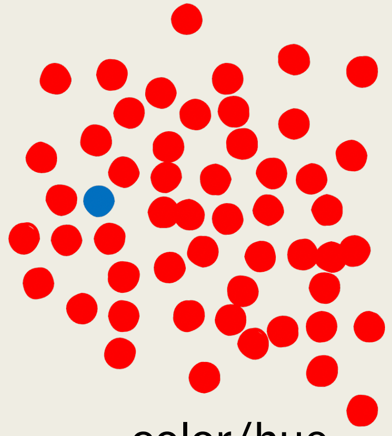
shape



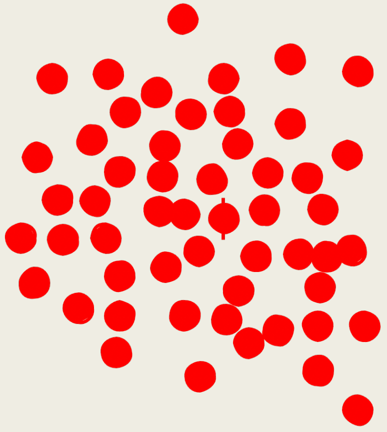
size



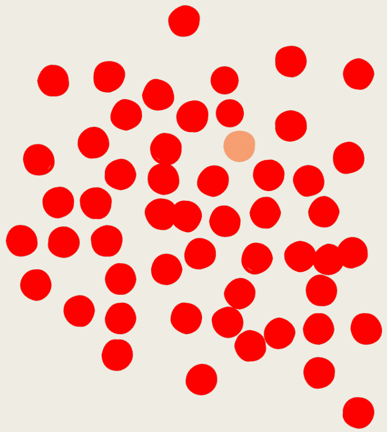
sharpness



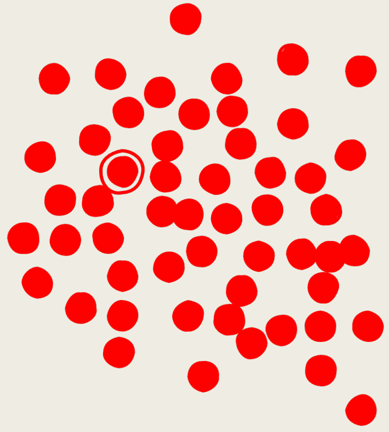
color/hue



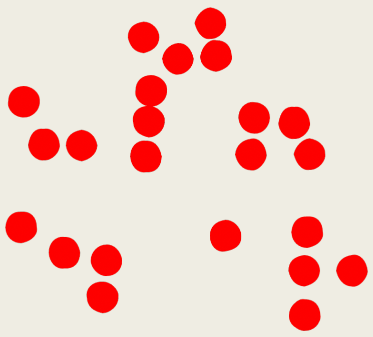
markings



intensity/value



enclosure



numerosity

[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]

PRE-ATTENTIVE ATTRIBUTES

How many 6’s are there on the next slide?

[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]

2869408609876

9348586748676

2967303986739

3967496749674

[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]

28**6**94086**6**09876**6**

93485586**6**7486**6**7**6**

29**6**730398**6**739

39**6**749**6**749674

2869408609876

9348586748676

2967303986739

3967496749674

[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]

2869408609876

9348586748676

2967303986739

3967496749674

[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]

2869408609876

934858**67**48**67**6

29**67**30398**67**39

39**67**49**67**49**67**4

[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]

2869408609876

9348586748676

2967303986739

3967496749674

DECLUTTERING

CLUTTER IS THE ENEMY!

- every element on a page adds **cognitive load**
- identify anything that isn't adding value and **remove**
- think of cognitive load as mental effort required to process information (lower is better)
- Tufte refers to the **data to ink ratio** – "the larger the share of a graphic's ink devoted to data, the better"
- in Resonate, Duarte refers to this as "**maximizing the signal-to-noise ratio**" where the signal is the information or the story we want to communicate.

DECLUTTERING

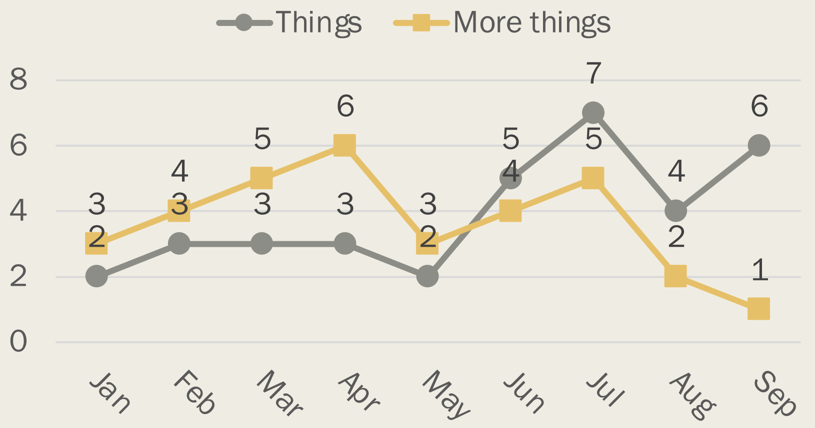
Use **Gestalt Principles** to organize/highlight data in a chart.

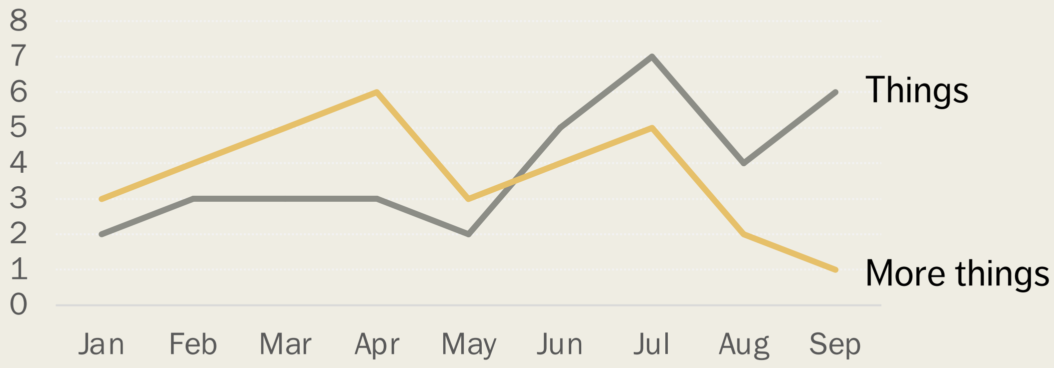
Align all the elements (graphs, text, lines, titles, etc.)

- DON'T rely on eye, use position boxes and values

Charts:

- remove border, gridlines, data markers
- clean up axis labels
- label data directly





DECLUTTERING

Use **consistent** font, font size, colour and alignment.

Don’t rotate text to anything other than 0 or 90 degrees.

Use **white space:**

- margins should remain free of text and visuals
- don’t stretch visuals to edge of page or too close to other visuals
- think of white space as a border

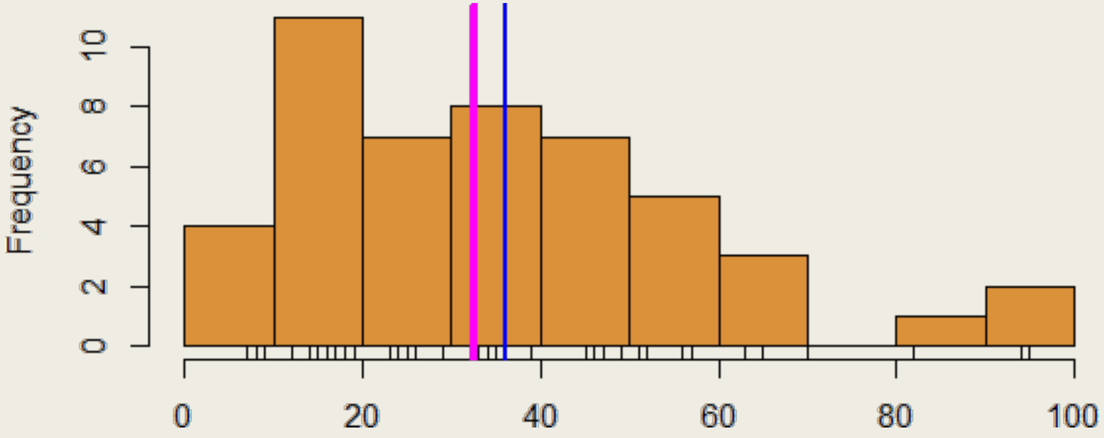
CHART SIZES

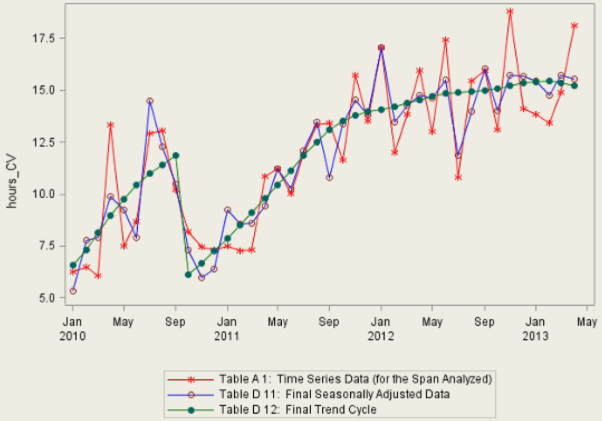
Assuming that the chart has been decluttered:

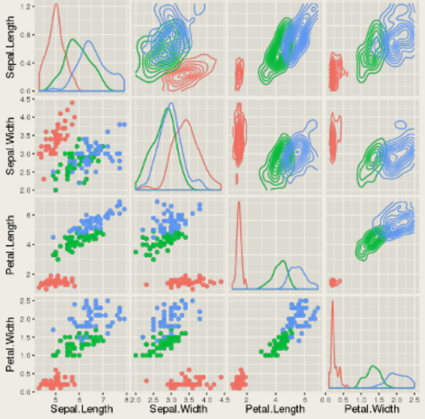
- things of equal importance size **similarly;**
- other things scale to **importance**.

As one rarely puts more than 3-4 charts on a page, there are limited size options.

Perennial exception: **geographical maps** may require more space.



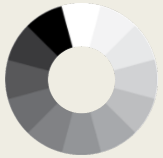




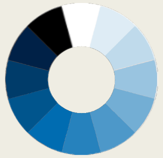
[http://www.deanenettles.com/webexamples/colorexamples]

COLOUR SCHEMES

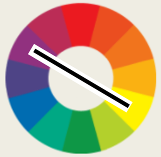
Achromatic



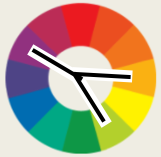
Monochromatic



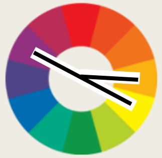
Complementary



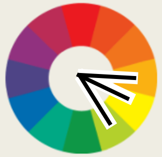
Split complementary



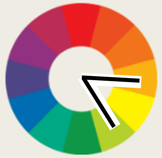
Split-Left/Right Complementary



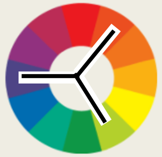
Analogous



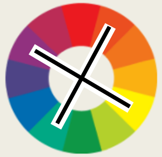
Colour Diad



Colour Triad

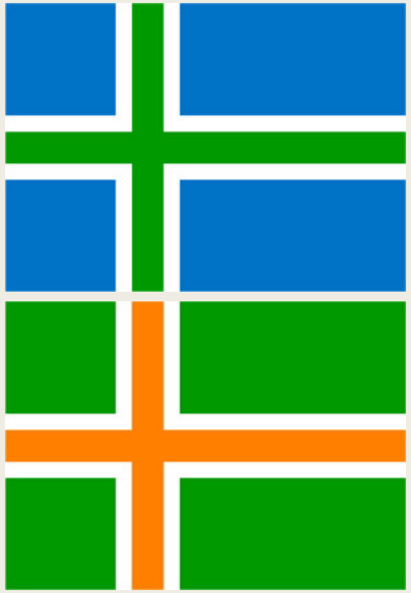


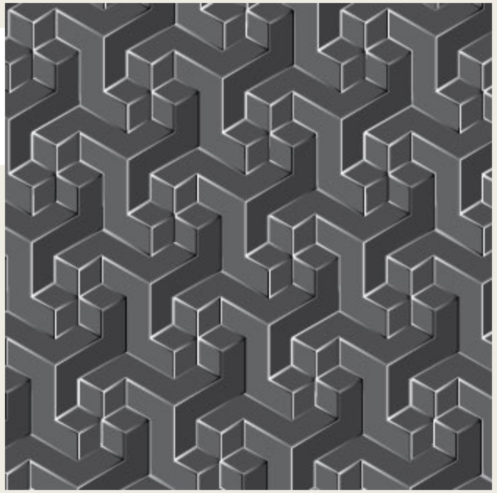
Colour Tetrad



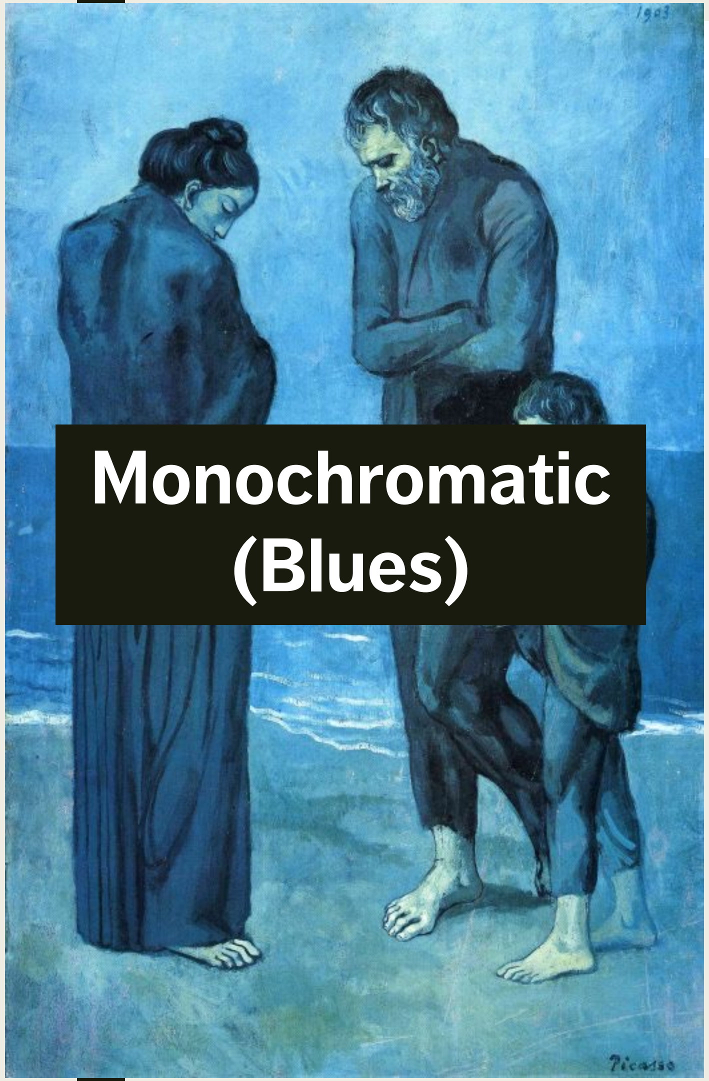
1903
Picasso

Claude Monet



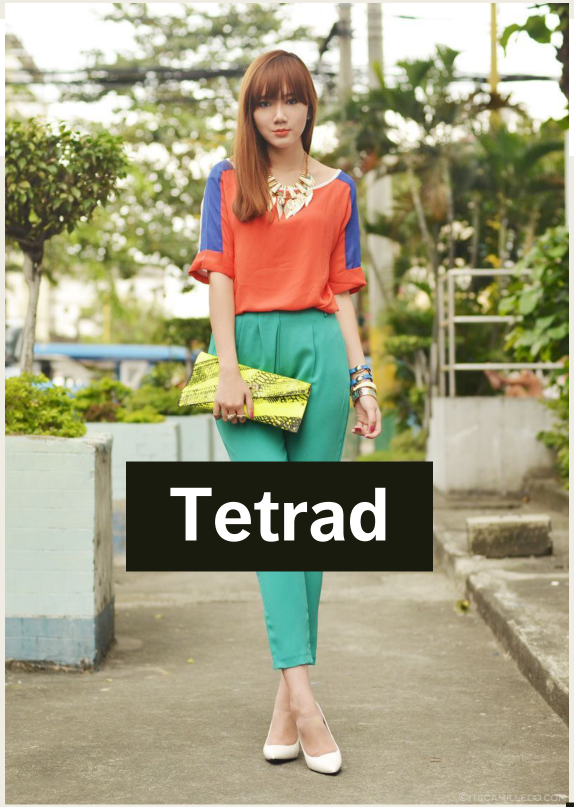


Can you identify the colour schemes underlying each of these images?

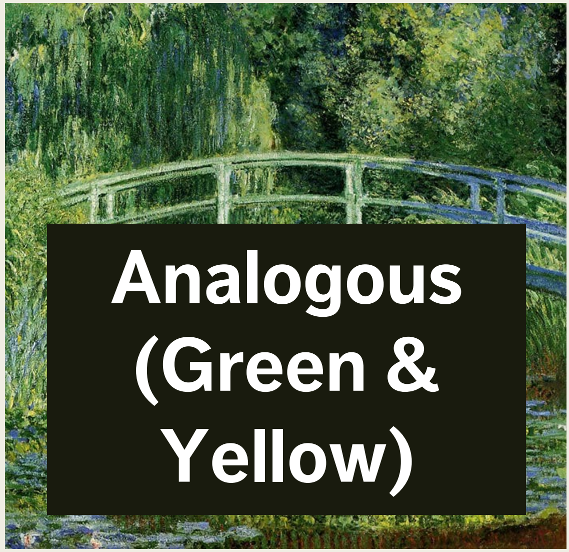


Monochromatic (Blues)

Can you identify the colour schemes underlying each of these images?

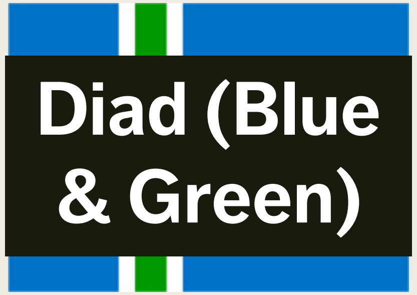


Tetrad



Analogous (Green & Yellow)

Split Complementary (Green, Orange & Blue)

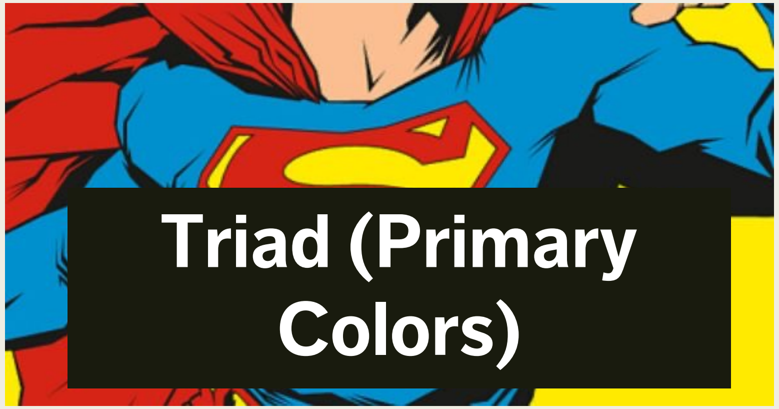


Diad (Blue & Green)

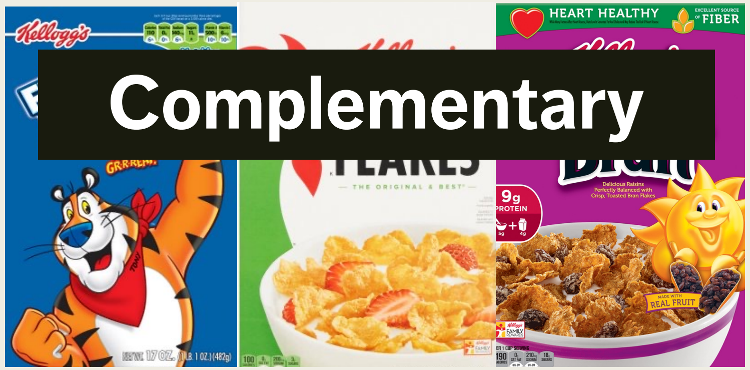
Diad (Green & Orange)

Diad (Red &Violet)

Achromatic

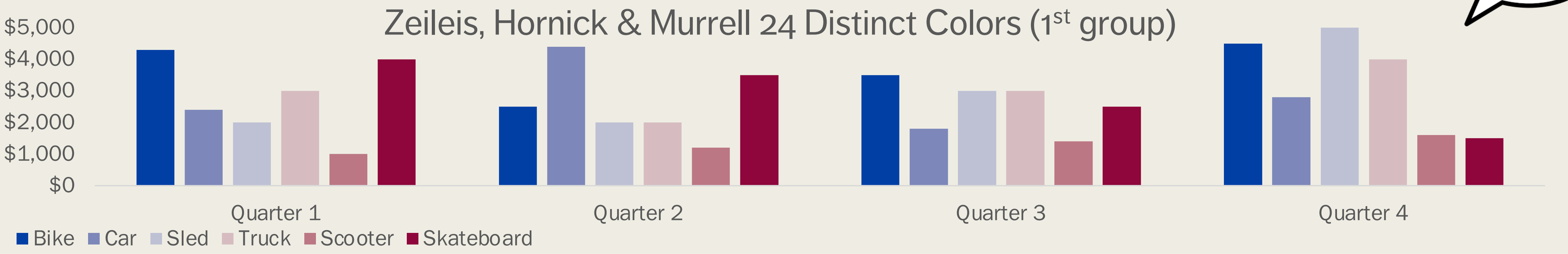


Triad (Primary Colors)

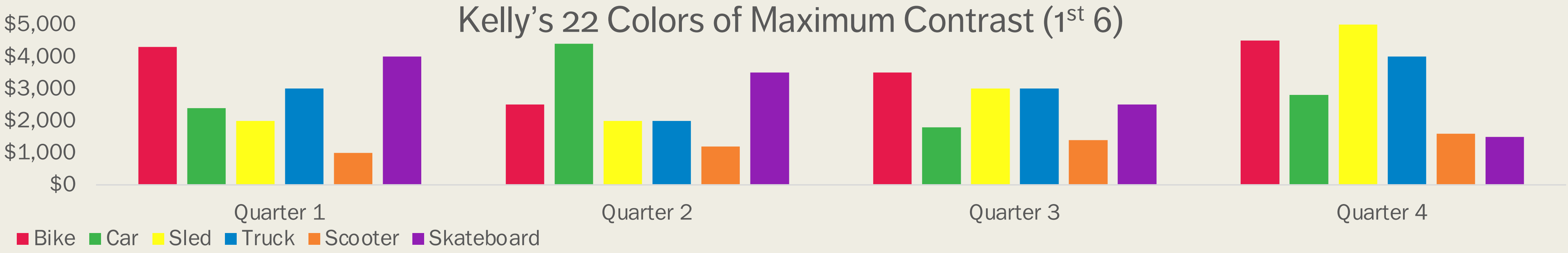


Complementary

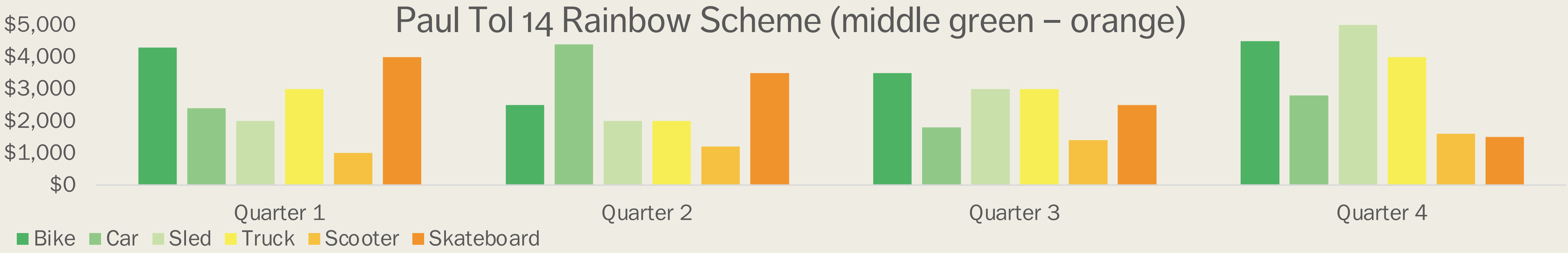
Zeileis, Hornick & Murrell 24 Distinct Colors (1st group)



Kelly's 22 Colors of Maximum Contrast (1st 6)



Paul Tol 14 Rainbow Scheme (middle green – orange)



COLOUR SCHEMES

When it comes to colour, **less is more**: use it sparingly (graphic designers are taught to "get it right, in black and white").

Based on the Gestalt Principles, **monochrome** schemes can be particularly effective.

When appropriate, pick scheme based on corporate identity (this maximizes buy in).

Create a template (and stick to it).

Upload images to see what charts look like in various flavours of colour-blindness:

- https://www.color-blindness.com/coblis-color-blindness-simulator (there are other tools)

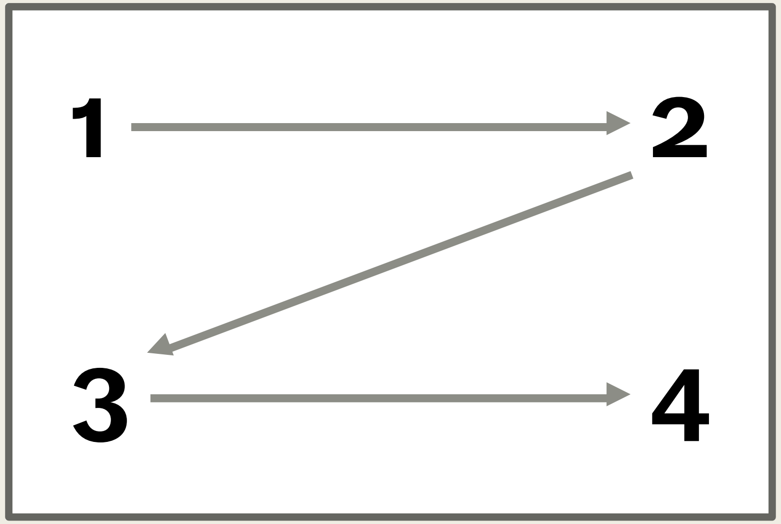
POSITION

How should the elements be placed in a chart or a dashboard?

In the West, most people start at the **top left** and zig- zag all the way to the **bottom right**.

Simple rule: don’t make people work too hard

- main message: top left/top right
- info in order of preference
- people concentrate less as they scan so get less complex as you move to bottom corner



DASHBOARDS

DATA VISUALIZATION BASICS

[Wexler, Shaffer, Cotgreave, *The Big Book of Dashboards*]

DASHBOARDS

A **dashboard** is any visual display of data used to monitor conditions and/or facilitate understanding.

Examples:

- interactive display that allows people to explore motor insurance claims by city, province, driver age, etc.
- PDF showing key audit metrics that gets e-mailed to a Department's DG on a weekly basis.
- wall-mounted screen that shows call centre statistics in real-time.
- mobile app that allow hospital administrators to review wait times on an hourly- and daily-basis for the current year and the previous year.

SOME QUESTIONS TO PONDER

In a car’s dashboard, a small number of **key indicators** (speed, gasoline level, lights, etc.) need to be understood **at a glance**. A dashboard design that does not take these two characteristics under consideration can have catastrophic consequences.

The following questions need to be answered prior to the dashboard being designed:

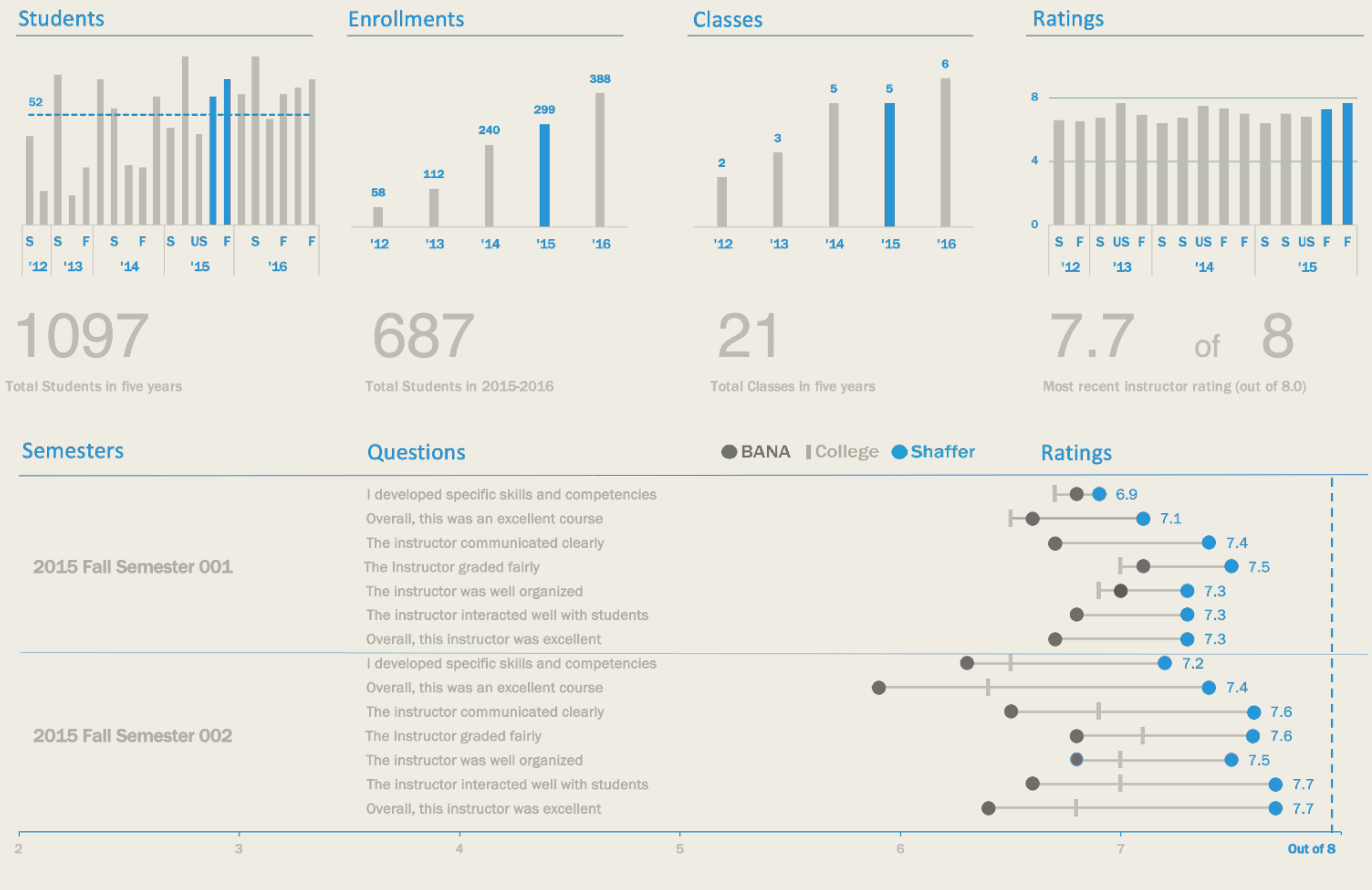
- Who is the dashboard’s **consumer**?
- What **story** does the dashboard tell?
- What data (categories) will be used?
- What will **appear** on the dashboard?
- How can the dashboard **help** the consumer?

Course Metrics

[https://bigbookofdashboards.com/dashboards.html]

Strengths:

- Easy-to-see key metrics
- Simple color scheme
- Potential to be static or interactive
- Both overview and details are clear



Course Metrics Dashboard created by Jeffrey A. Shaffer. Data from University of Cincinnati Course Evaluations. Blue indicates the 2 most recent rating periods.

DASHBOARD EVALUATION

There are no perfect dashboards – no collection of charts will ever suit everyone who encounters it.

All dashboards should be **truthful** and **functional**, but dashboards that are also **elegant** (delightful, enjoyable) will take you further.

All dashboards are **incomplete**. Good dashboards will still lead to dead ends, but they should allow users to ask: “Why? What is the root cause of a problem?”

Tools: Excel, Power BI, Tableau, R + Shiny, Geckoboard, Matillion, etc.

EXERCISE

Consider the following dashboards.

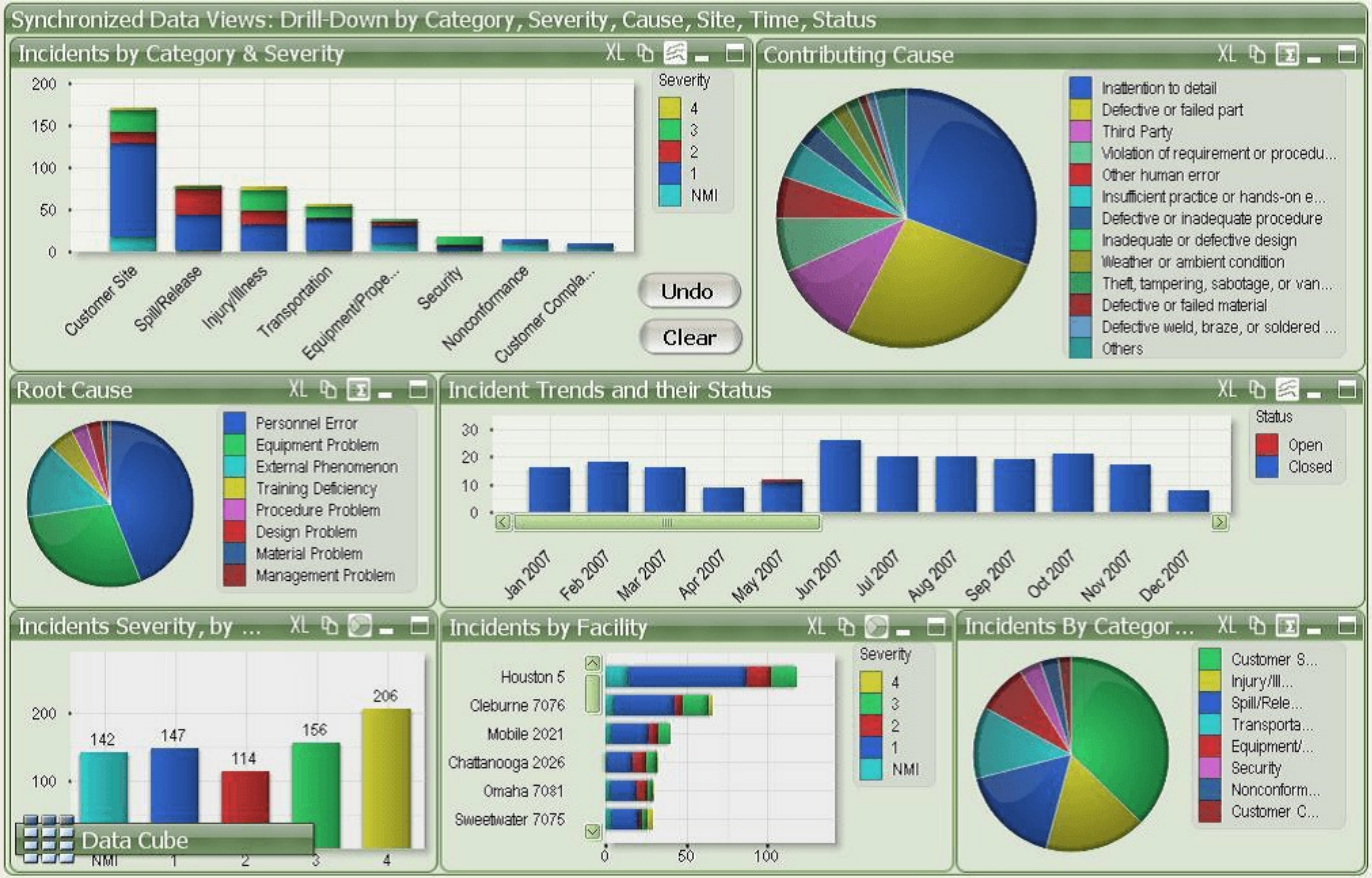
Can you figure out, at a glance, who their audience is?

What are their strengths?

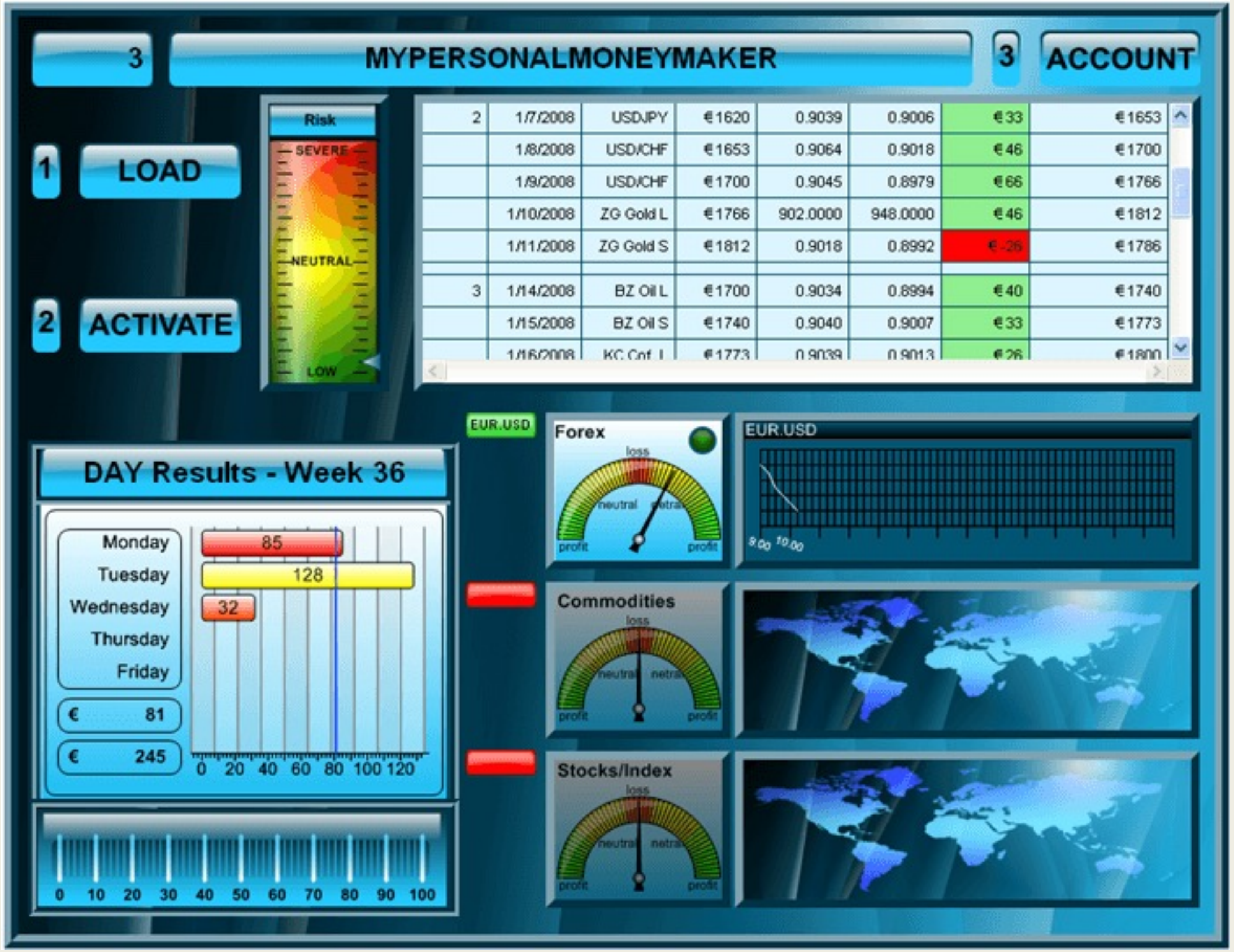
What are their limitations?

How would you improve them?

[https://www.matillion.com/wp-content/uploads/2014/11/qlikview-poor-use-of-dashboard-software.png]



[https://www.geckoboard.com/assets/2-terrible-dashboard-example-min.png]



2	1/7/2008	USDJPY	€1620	0.9039	0.9006	€33	€1653
	1/8/2008	USD/CHF	€1653	0.9064	0.9018	€46	€1700
	1/9/2008	USD/CHF	€1700	0.9045	0.8979	€66	€1766
	1/10/2008	ZG Gold L	€1766	902.0000	948.0000	€46	€1812
	1/11/2008	ZG Gold S	€1812	0.9018	0.8992	€ -26	€1786
3	1/14/2008	BZ Oil L	€1700	0.9034	0.8994	€40	€1740
	1/15/2008	BZ Oil S	€1740	0.9040	0.9007	€33	€1773
	1/16/2008	KC Cof L	€1773	0.9039	0.9013	€26	€1800