Principles of Data Visualization for Exploratory Data Analysis

Renee M. P. Teate

UVA SYS 6023
Cognitive Systems Engineering
Spring 2015

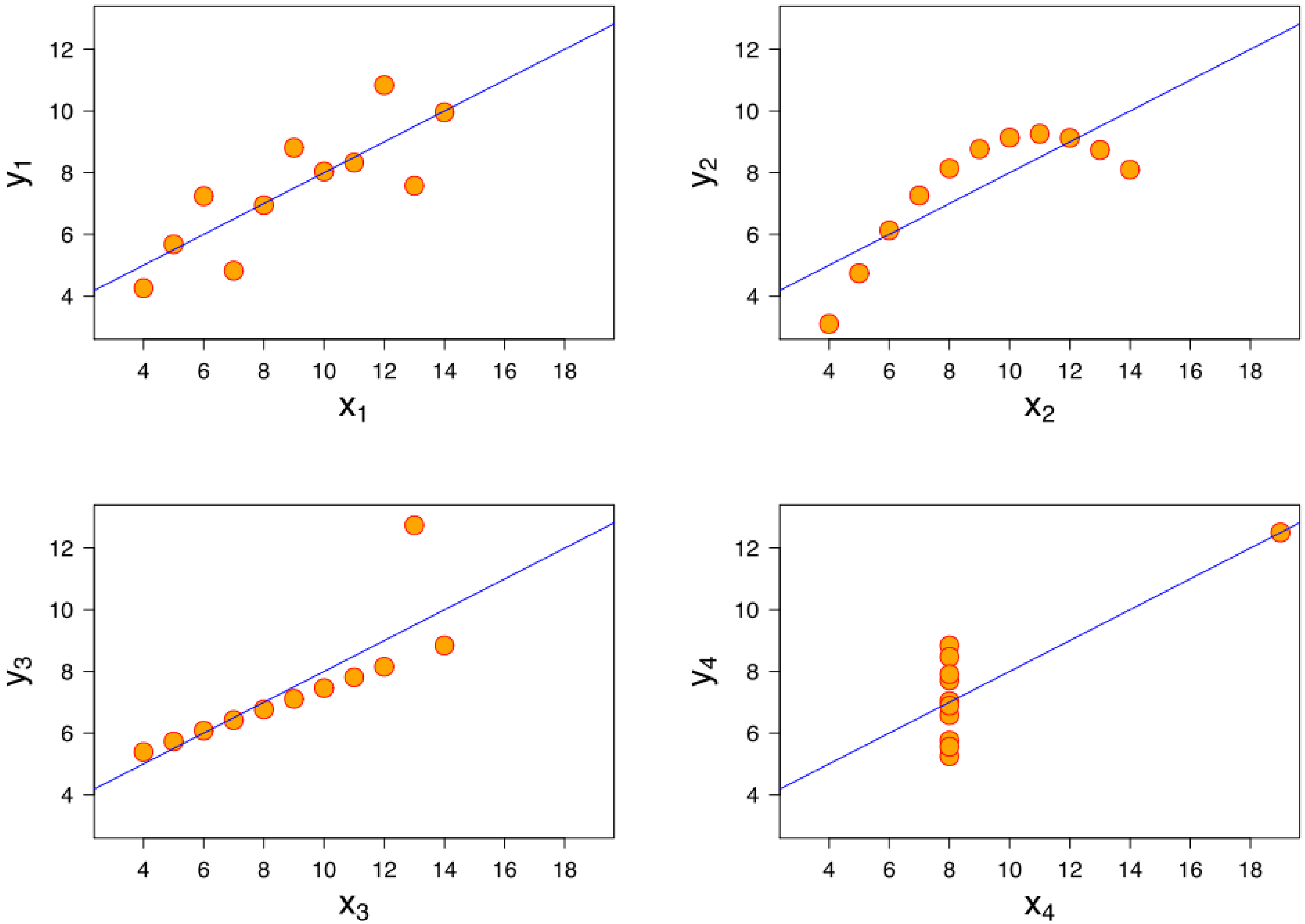
What is Data Visualization?

- Quantitative data presented in visual form[1]
  - Supports exploration, examination, and communication of information[1]
  - Common characteristics: computer-supported, interactive, visual representation, abstract, amplifies cognition[1]
- 2 objectives:
  1. Analysis
  2. Communication[2]

Why Visualize Data?

- Humans generally poor at gaining insight from data in numerical form[3]
- Close relationship between vision & cognition[1]
- Allows you to explore and make sense of data, and communicate information[5]
- Make patterns, trends, exceptions visible and understandable[1]
- Extend capacity of memory – puts in front of eyes what we couldn’t otherwise hold in mind[1]
- Especially useful when little known about data and analysis goals are vague[6]
- Can help with hypothesis generation[6]

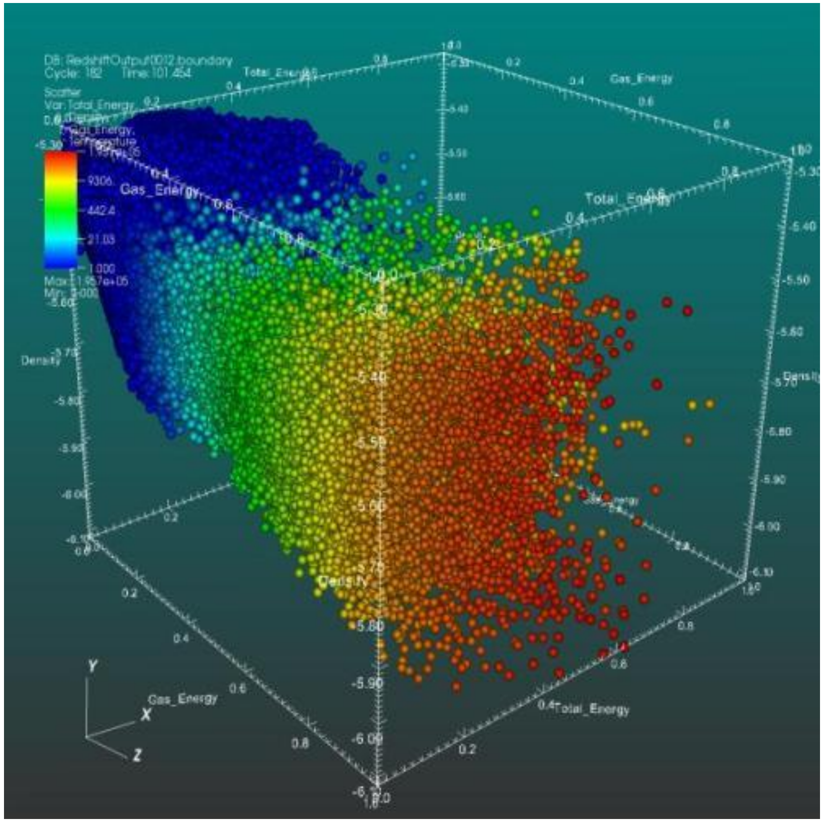
Anscombe's quartet



"Anscombe's quartet 3" by Schutz via Wikimedia Commons

"One great virtue of good graphical representation is that it can serve to display clearly and effectively a message carried by quantities whose calculation or observation is far from simple."

– John W. Tukey[1]



"[Scatter plot]" by UCRL via Wikimedia Commons

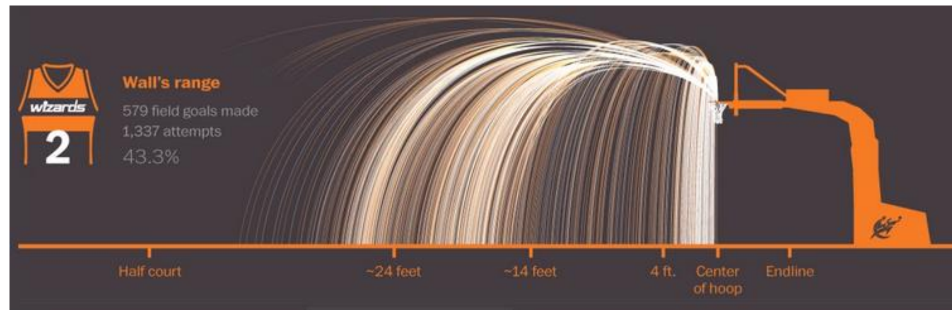
Is this a "good" data visualization?

(more on this later)

Illustration vs Visualization

Data Illustration:

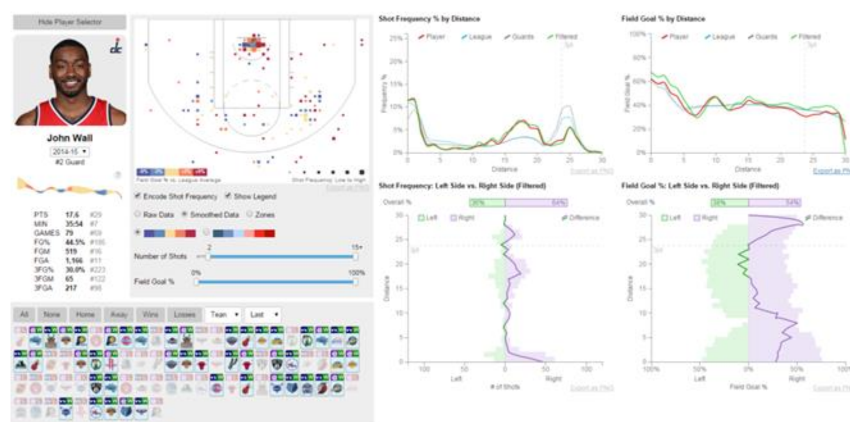
- To impress, inspire awe, make people wonder[7]
  - Memorable & engaging vs comprehensible[8]



"Wizards Shooting Stars" Washington Post via [FlowingData]()

Data Visualization:

- To inform[7]
  - Explore, Make sense of, and Communicate[5]
  - Optimal for:
    - Seeing big picture
    - Rapidly comparing values
    - Seeing patterns among values
    - Comparing patterns across multiple sets[5]



"Buckets" by Peter Beshai via [FlowingData]()

“wind map” - http://hint.fm/wind

WolframAlpha computational knowledge engine

wind 22801

Examples Random

WolframAlpha results for "wind 22801"

Input interpretation:

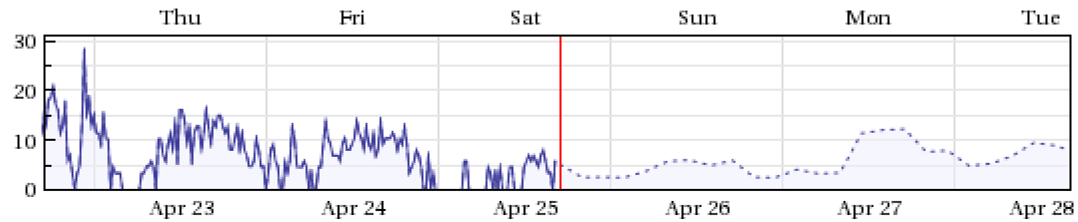
wind speed	ZIP code 22801

Result: Show metric

6 mph (miles per hour) 180° S

(47 minutes ago)

History & forecast: Current week ▼ Show metric More



low: 0 mph Sat, Apr 25, 4:00pm, ...	average: 6 mph	high: 28 mph Wed, Apr 22, 10:30pm

Weather station information: Show metric More

name	KSHD (Shenandoah Valley Regional Airport)
relative position	10 mi S (from ZIP code 22801)
relative elevation	(comparable to ZIP code 22801)
local time	5:01:34 pm EDT | Saturday, April 25, 2015
local sunlight	sun is above the horizon azimuth: 259° (W) | altitude: 34°

Units »

Satellite image »

What is Exploratory Data Analysis (EDA)?

"Seeing what the data can tell us"

- Initial examination of a dataset:
  - Determine data types, summary statistics
  - Assess your assumptions about the data
  - Start forming hypothesis about phenomenon you observe[9]
  - Question everything; Ask "why" often
  - Explore outliers[10]
- Supports selection of tools & techniques[9]
- Can provide basis for additional data collection[9]
- Verify what you know, expose what you don't[10]

Combining the concepts: Visual Exploratory Data Analysis

- For this study, I searched for information related to visuals that:
  - Are most helpful to analysts during this exploratory stage
  - Can be generated quickly
  - Are for analysis, not necessarily communication (i.e. don't have to follow all "best practices" for accessibility, information sharing, or publication at this point)
  - Take advantage of human visual perceptual strengths

"Information Seeking Mantra"

Overview first, zoom and filter, then details-on-demand[11]

A look at two Basic Data Visualization Types for EDA: **Bar Graphs & Line Graphs**

Bar Graphs

- Imply individual values[12]
- Accurately show fixed intervals[13]
- Used to plot categorical vs quantitative data
- Can be horizontal or vertical
  - Should always use vertical when categories represent time periods
  - Horizontal when long categorical labels needed
- Can be used to show distribution as Histogram where categories are buckets of the same interval size

DESIGN PRINCIPLES

- Axis must start at 0 to support comparing values, otherwise misleading
- Distance between bars, width of bars have no quantitative meaning
- Consider how bars are grouped
- Use light colors if needed

[All unmarked bullets on slide are from reference 4]

Line Graphs

- Imply transitions[12]
- Looks continuous[13]
- Avoid for nominal comparisons or rankings
- Can connect points in time series if intervals consistent
- Show values, changes, deviations, distributions
- Can be overlaid on other graph types to show trends or reference values

DESIGN PRINCIPLES

- Aspect Ratio is important
- Ensure multiple lines are visually distinct, can use medium colors
- Only include points when viewer needs to compare instances across lines
- Typically linear scale, but Log scale allows comparison of rates of change
- Label lines directly if possible instead of using legend

Bar Graphs

- Imply individual values [12]
- Accurately show fixed intervals[13]
- Used to plot categorical vs quantitative data
- Can be horizontal or vertical
  - Should always use vertical when categories represent time periods
  - Horizontal when long categorical labels needed
- Can be used to show distribution as Histogram where categories are buckets of the same interval size

DESIGN PRINCIPLES

- Axis must start at 0 to support comparing values, otherwise misleading
- Distance between bars, width of bars have no quantitative meaning
- Consider how bars are grouped
- Use light colors if needed

[All unmarked bullets on slide are from reference 4]

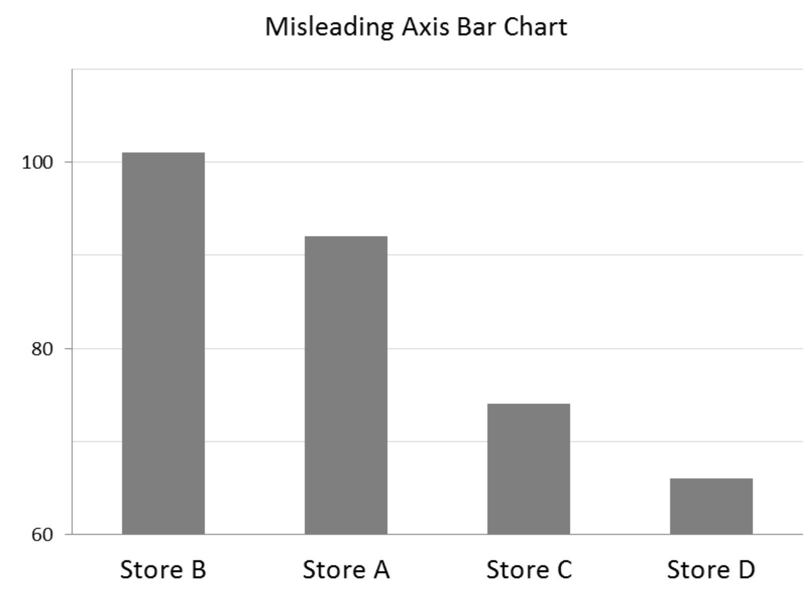
Line Graphs

- Imply transitions [12]
- Looks continuous[13]
- Avoid for nominal comparisons or rankings
- Can connect points in time series if intervals consistent
- Show values, changes, deviations, distributions
- Can be overlaid on other graph types to show trends or reference values

DESIGN PRINCIPLES

- Aspect Ratio is important
- Ensure multiple lines are visually distinct, can use medium colors
- Only include points when viewer needs to compare instances across lines
- Typically linear scale, but Log scale allows comparison of rates of change
- Label lines directly if possible instead of using legend

Example Perception-Based Design Principle

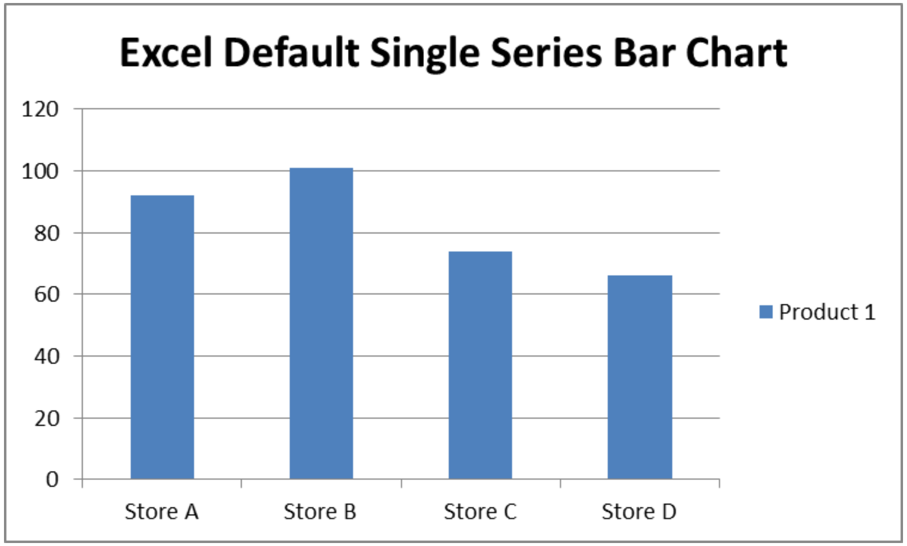


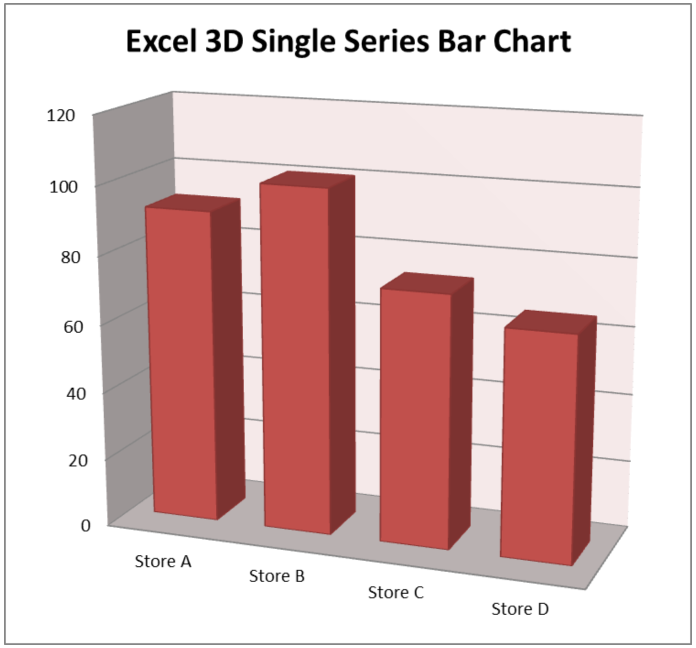
The axis on a bar graph must start at 0, because we perceive the differences between the bar heights as proportional.
(i.e. a bar twice as tall represents a value twice as large) [4]

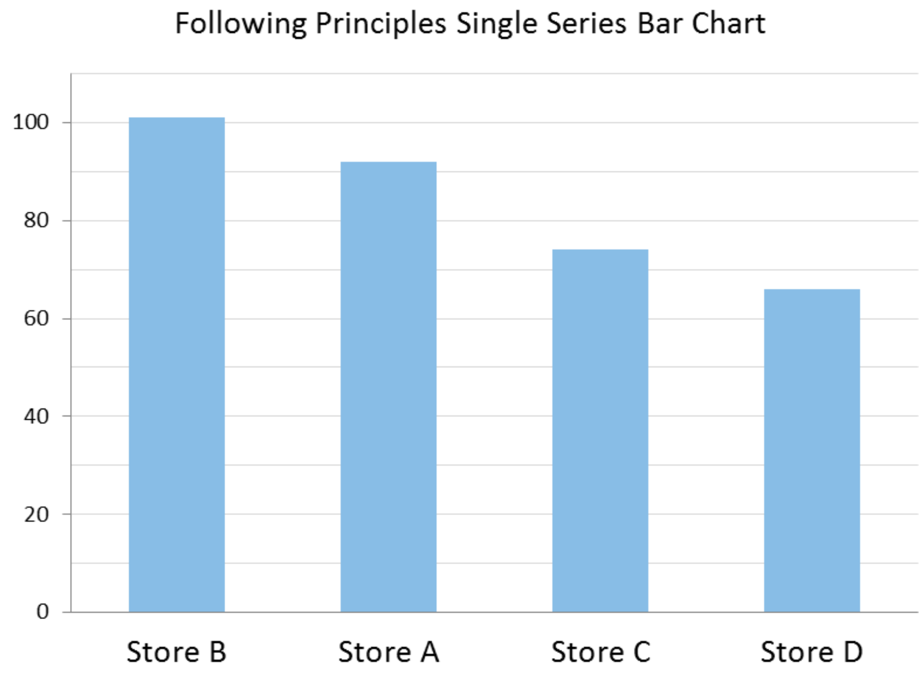
Can you gain much insight from this set of data without a visual?

Product	Store A	Store B	Store C	Store D
Product 1	92	101	74	66
Product 2	28	90	52	75
Product 3	15	21	7	-10

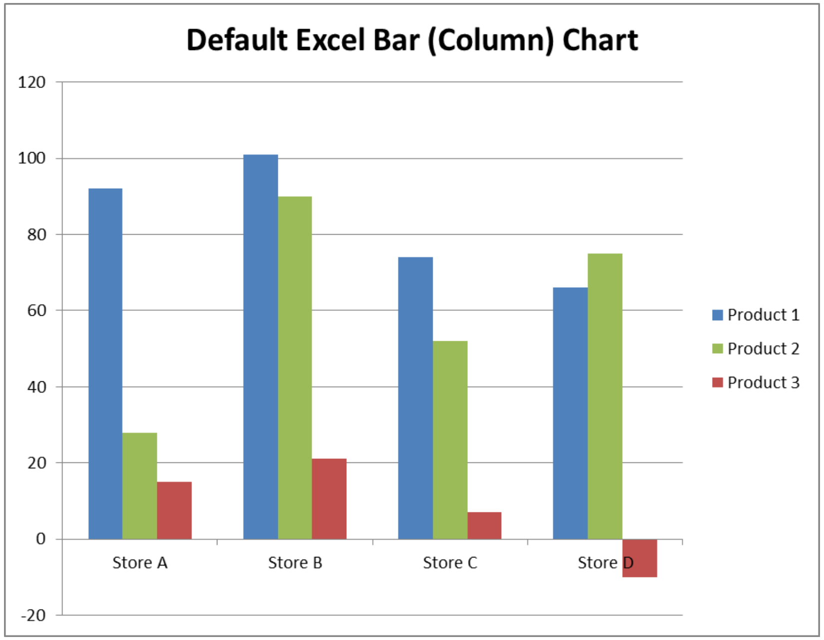
Let’s create some graphs.

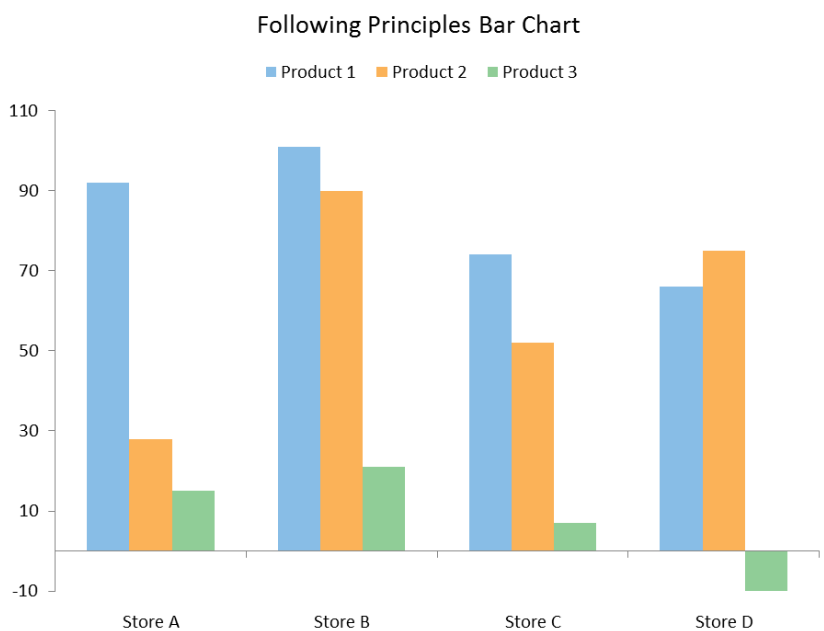






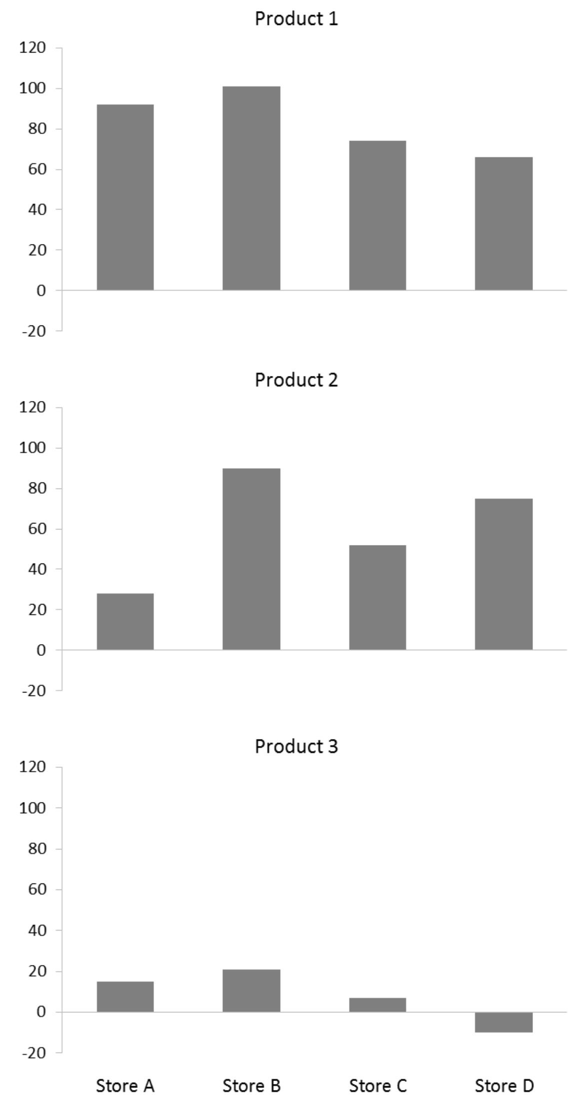
Design principles from "Show Me the Numbers" by Stephen Few[4]



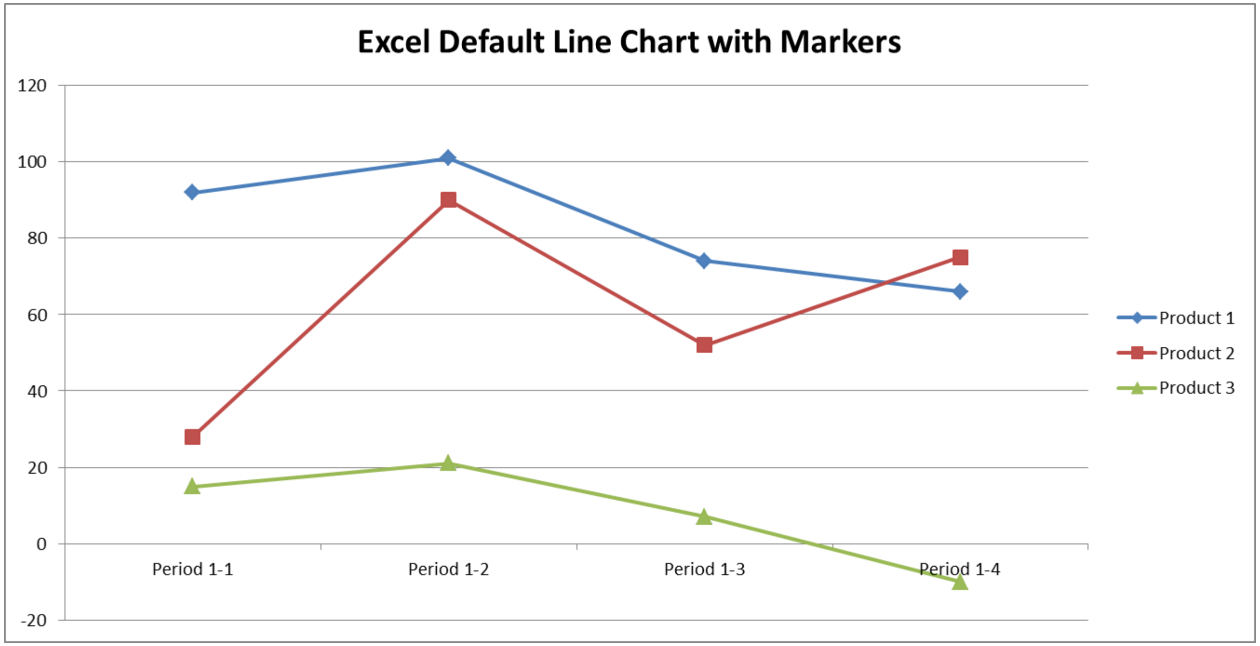


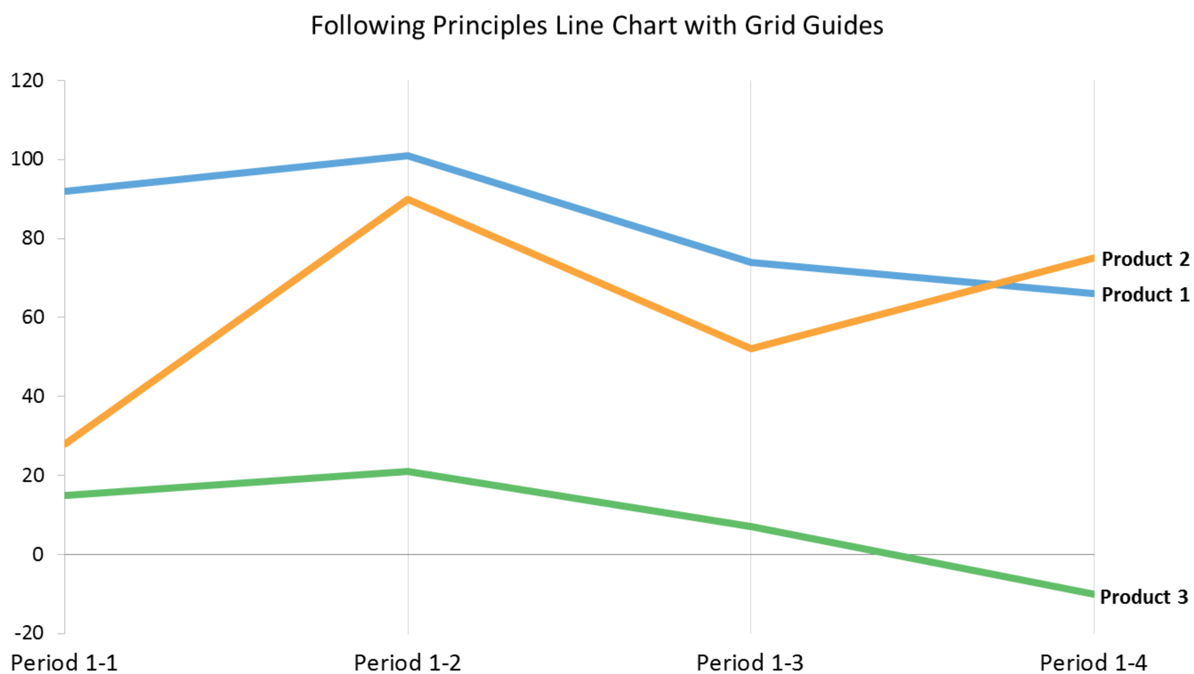
Another option: Small Multiples

(makes a bigger difference with more series)



Design principles from "Show Me the Numbers" by Stephen Few[4]





Design principles from "Show Me the Numbers" by Stephen Few[4]

Perception of **Multidimensional Data Visualizations**

What happens when we need to encode more than 3 attributes on a visual?

Like month, sales in dollars, sales person, office location…

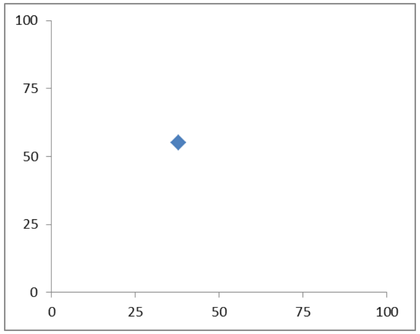
Bertin's Image Theory[3]

We can only perceive 3 variables (2 planar and 1 retinal) "efficiently".
Efficient = preattentive, without additional eye motion or attention required.

PLANAR

Spatial dimension 1

Spatial dimension 2



RETINAL

Texture

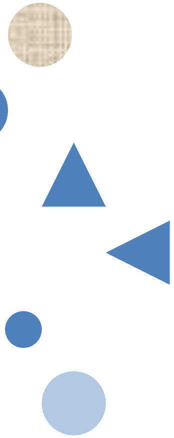
Color

Shape

Orientation

Size

Brightness



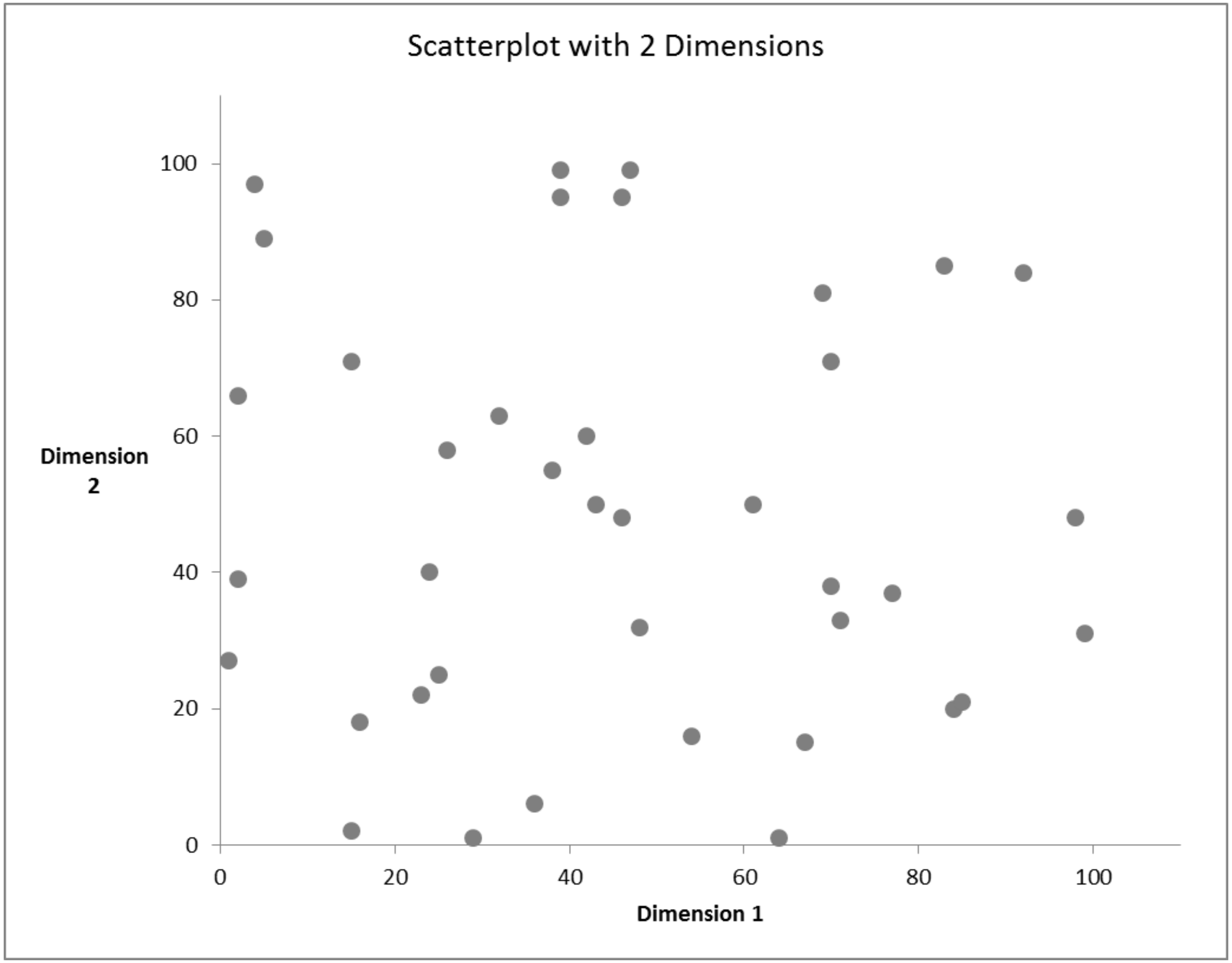
This means that humans can not effectively visualize 4 dimensions using a graphical representation on a 2-dimensional display (screen or paper).

Bertin’s Image Theory

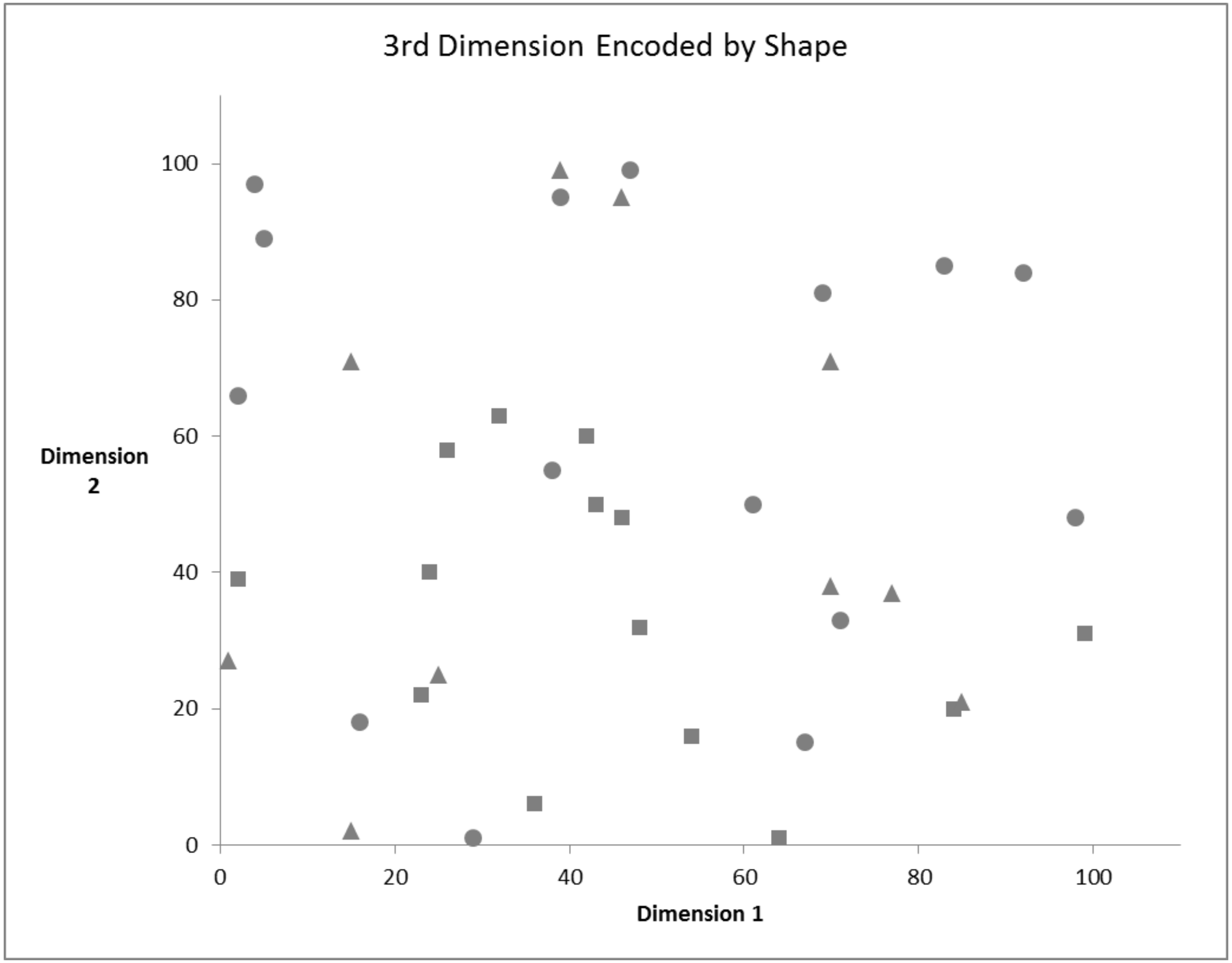
Table I: Original Bertin				
	Associative	**Selective**	**Ordered**	**Quantitative**
Planar	Yes	Yes	Yes	Yes
Size		Yes	Yes	Yes
Brightness		Yes	Yes	
Texture	Yes	Yes	Yes	
Color	Yes	Yes		
Orientation	Yes	Yes		
Shape	Yes			

Skipping definitions of the columns in interest of time, but as an example :
Shape is neither ordered nor quantitative because it can’t be scaled for magnitude. (Does a triangle represent a larger value than a square?)

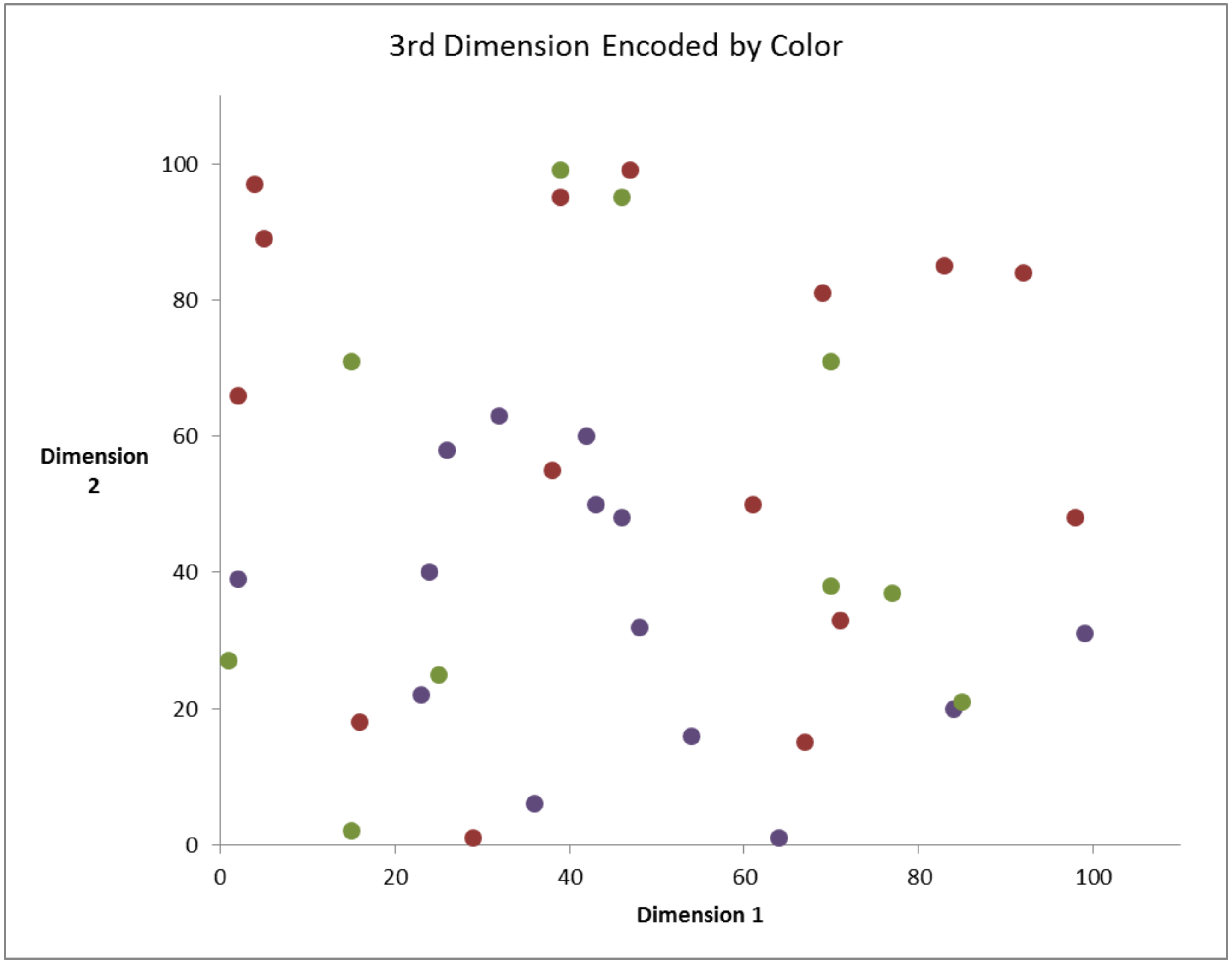
Bertin says that **failure to match the component and the visual “level” (type of scale) is the single major source of error in design of visualizations**. [3]



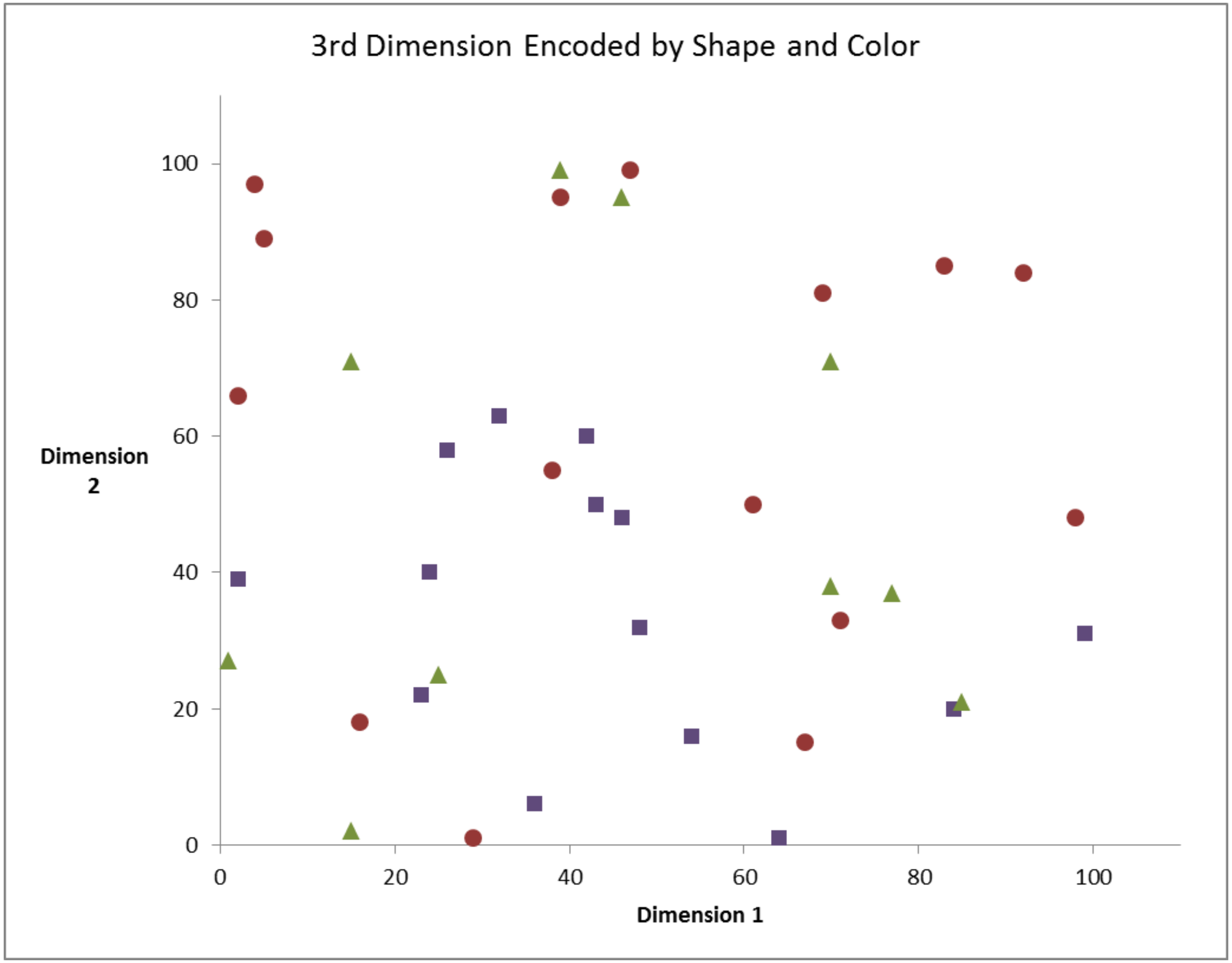
2 Correspondences – X location, Y location
2 Spatial, 0 Retinal



3 Correspondences – X location, Y location, Shape
2 Spatial, 1 Retinal

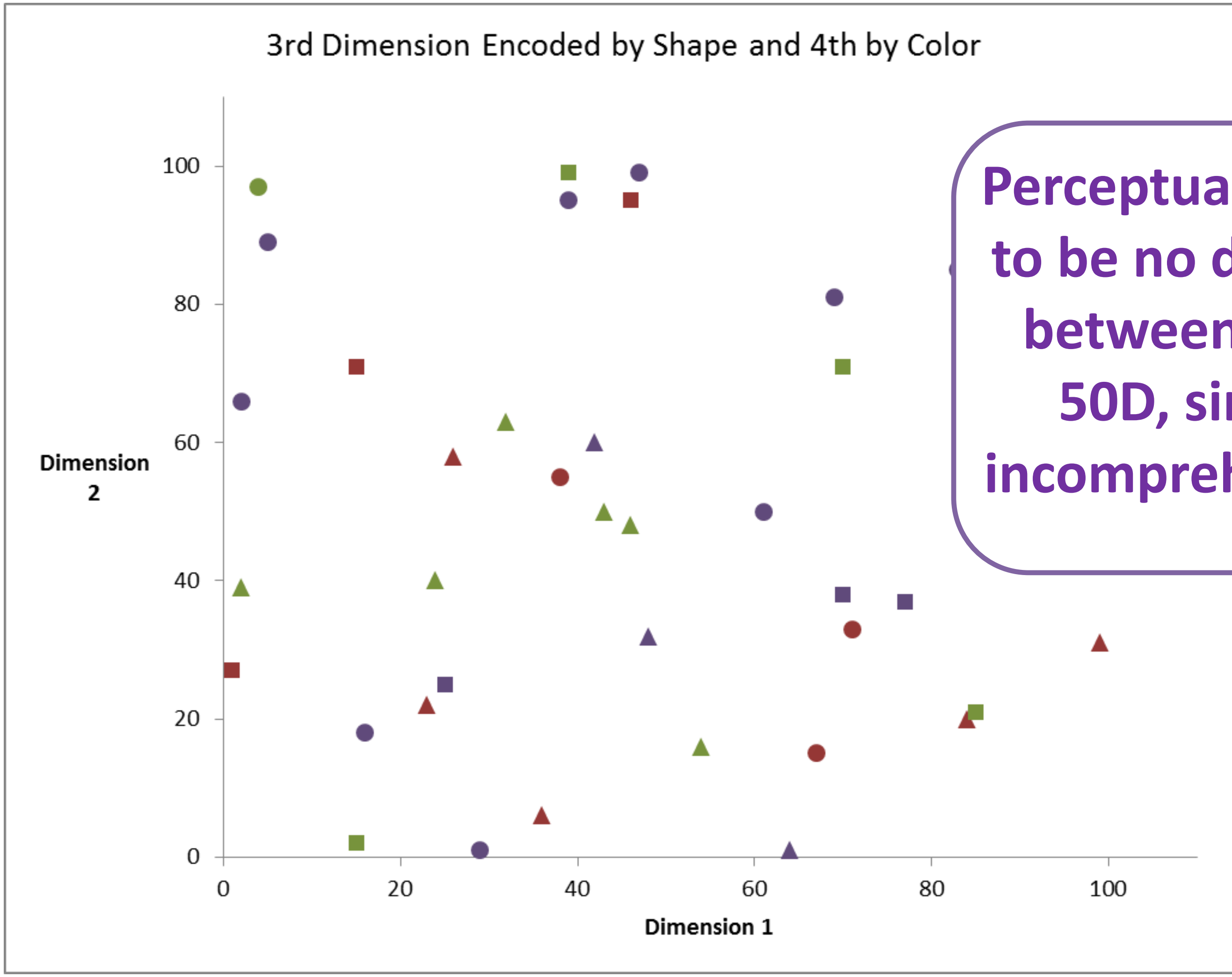


3 Correspondences – X location, Y location, Color
2 Spatial, 1 Retinal



4 Correspondences – X location, Y location, Color and Shape
2 Spatial, 2 Retinal (working as 1, encoding same dimension)

Human vision appears to only be able to differentiate 3 dimensions "efficiently".



Perceptually, seems to be no difference between 5D and 50D, similarly incomprehensible[14]

4 Correspondences – X location, Y location, Color, Shape
2 Spatial, 2 Retinal (encoding different dimensions)

A few notes on color perception

- Colorblindness an issue in graphs for communication, but not as much for analysis (unless you're building a tool for others)
- Rainbow scales are not good perceptually
  - We can visually order small ranges of hue, but not across entire spectrum
- Brightness can be used for ordering values
  - Each "level" must be perceptibly different
  - Doesn't linearly map to quantity[3]
- Colors suggested for heatmaps:
  - Blue to gray to Red[1]

- Brewer palettes (colorbrewer.org) provide a range of palettes based on HSV model which make life easier for us….

Avoid the use of hue to encode quantitative variables

Quantitative encoding e.g. heat maps

Two-sided quantitative encodings

Fig. Courtesy of M. Krzwinski,

Note that ROYGBIV Rainbow is not Quantitative or Ordered

Brewer palette slides from Principles of Information Visualization Tutorial – Jessie Kennedy[2]

Colour Blindness

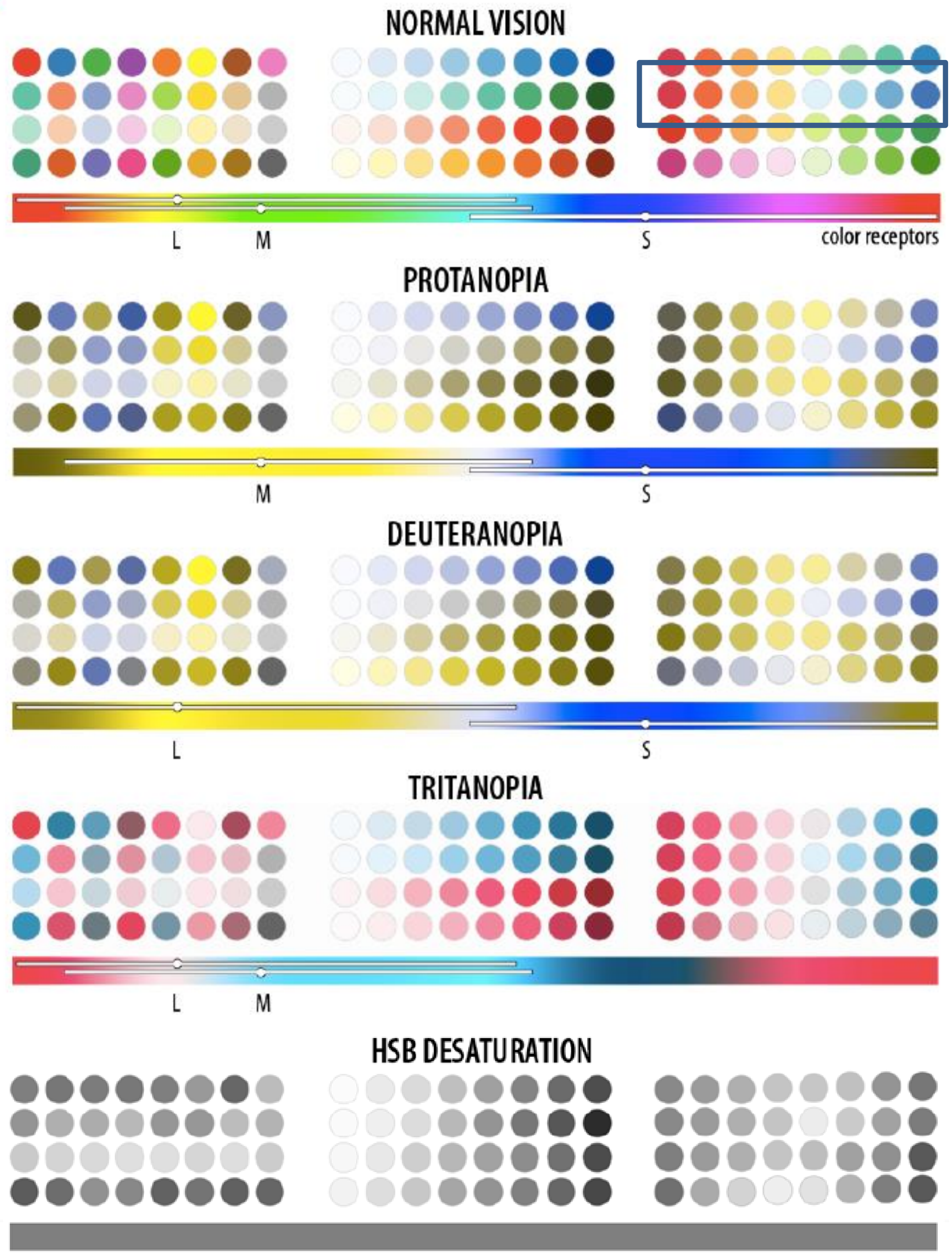
8% males of USA descent

Red-green

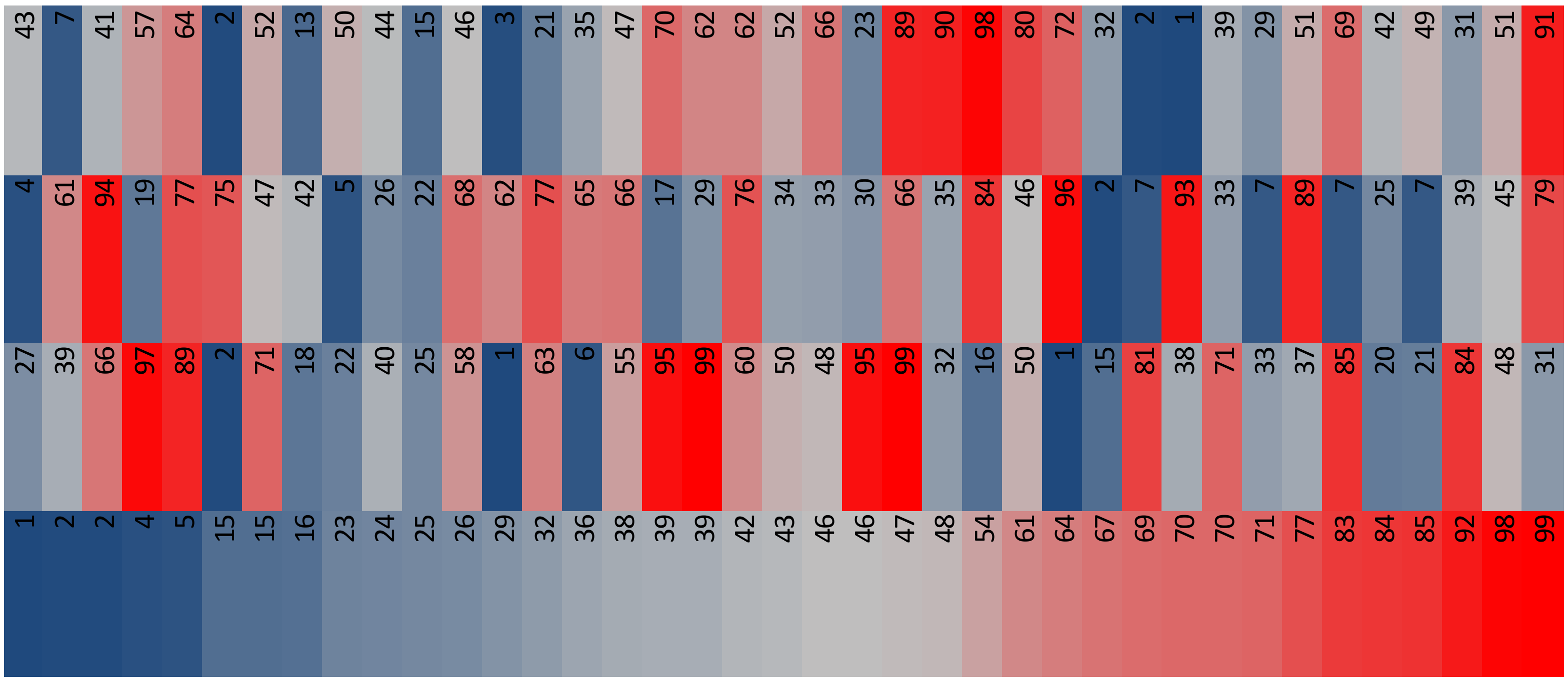
Red-green

Blue-yellow

Fig. Courtesy of M Krzywinski



Brewer palette slides from Principles of Information Visualization Tutorial – Jessie Kennedy[2] “**high**” red through “**neutral**” gray through “**low**” blue heatmap



Why think about all of this?

As an analyst, you should follow
as many **perception-based design principles**
as possible when making graphs during
Exploratory Data Analysis.

Good visualizations can help you
make sense of the data,
and spot patterns, trends, and exceptions
with the **least effort**.

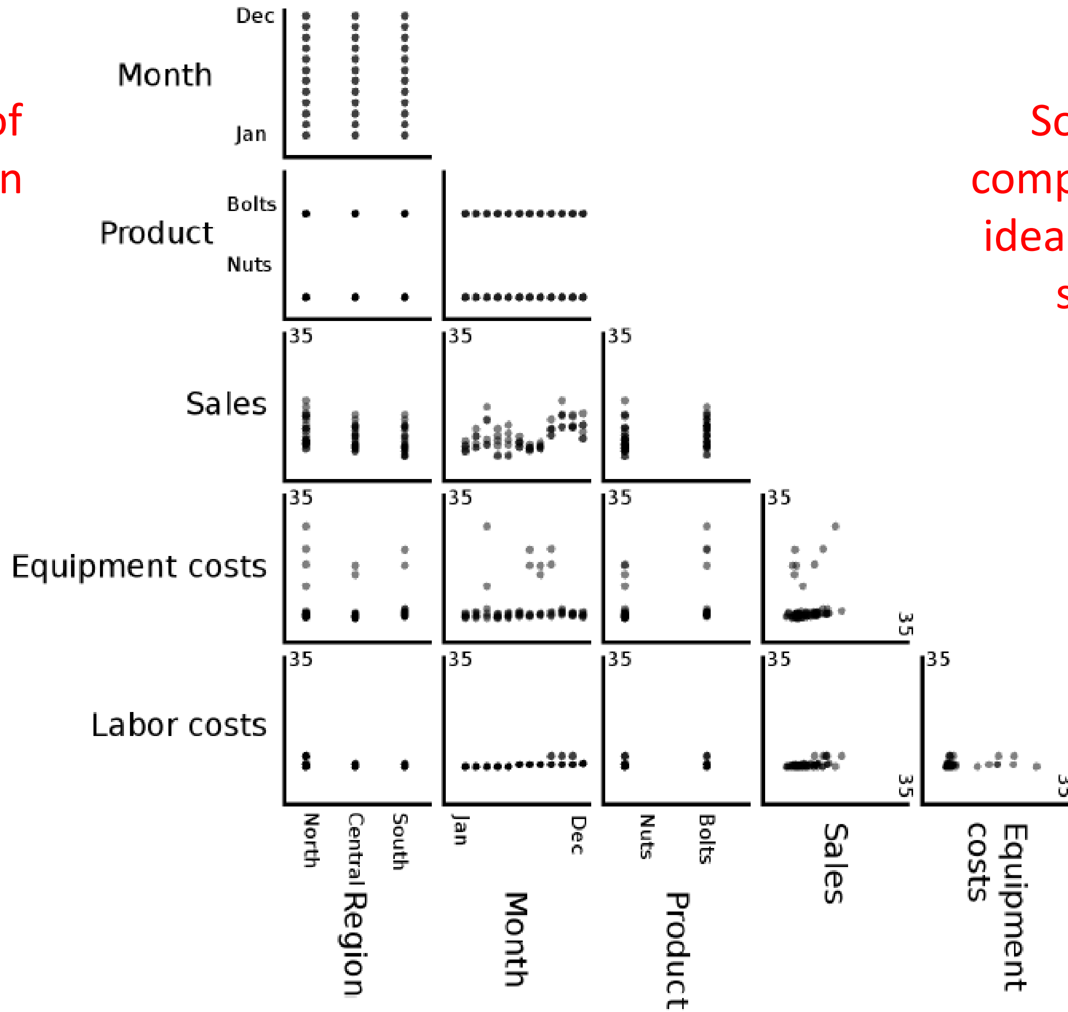
You can ensure you will **spot things
that would otherwise be hidden**
or difficult to perceive.

Other Techniques to Consider

- These are not necessarily “quick and easy” to create using common software, but there are tools available to take advantage of other strengths of human perception during EDA
  - **Scatterplot Matrix or GPLOM**
    - A form of “small multiples”
    - Allow many comparisons in one view
  - **Animation**
    - We’re good at spotting motion
    - Can help understand changes in multiple dimensions over time

Scatterplot Matrix

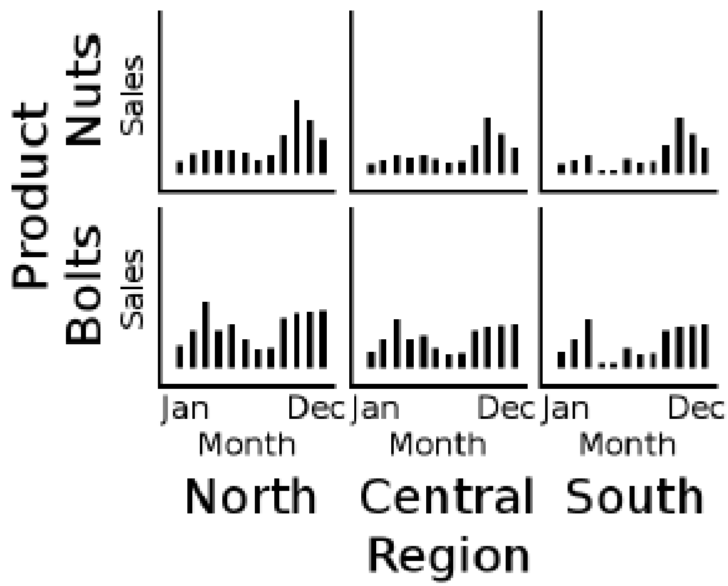
Allows comparison of every data dimension vs every other data dimension



Some of these comparisons are not ideally displayed as scatterplots

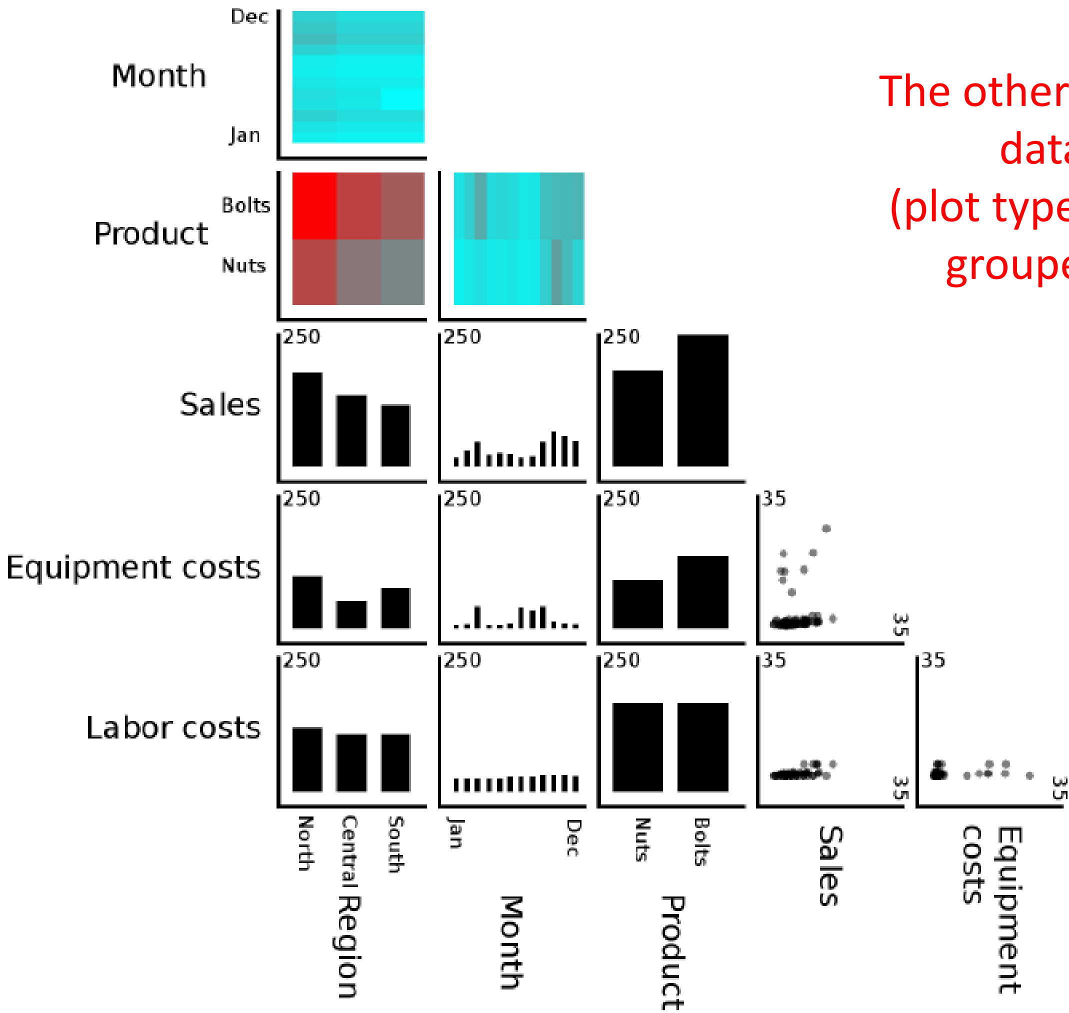
Images from GPLOM paper – Im, McGuffin, Leung[15]

Can split those out into dimensionally-aligned bar charts



Images from GPLOM paper – Im, McGuffin, Leung[15]

New design: Generalized Plot Matrix (GPLOM)



The other charts now show aggregated data for easier comparison (plot types automatically selected and grouped together in the display)

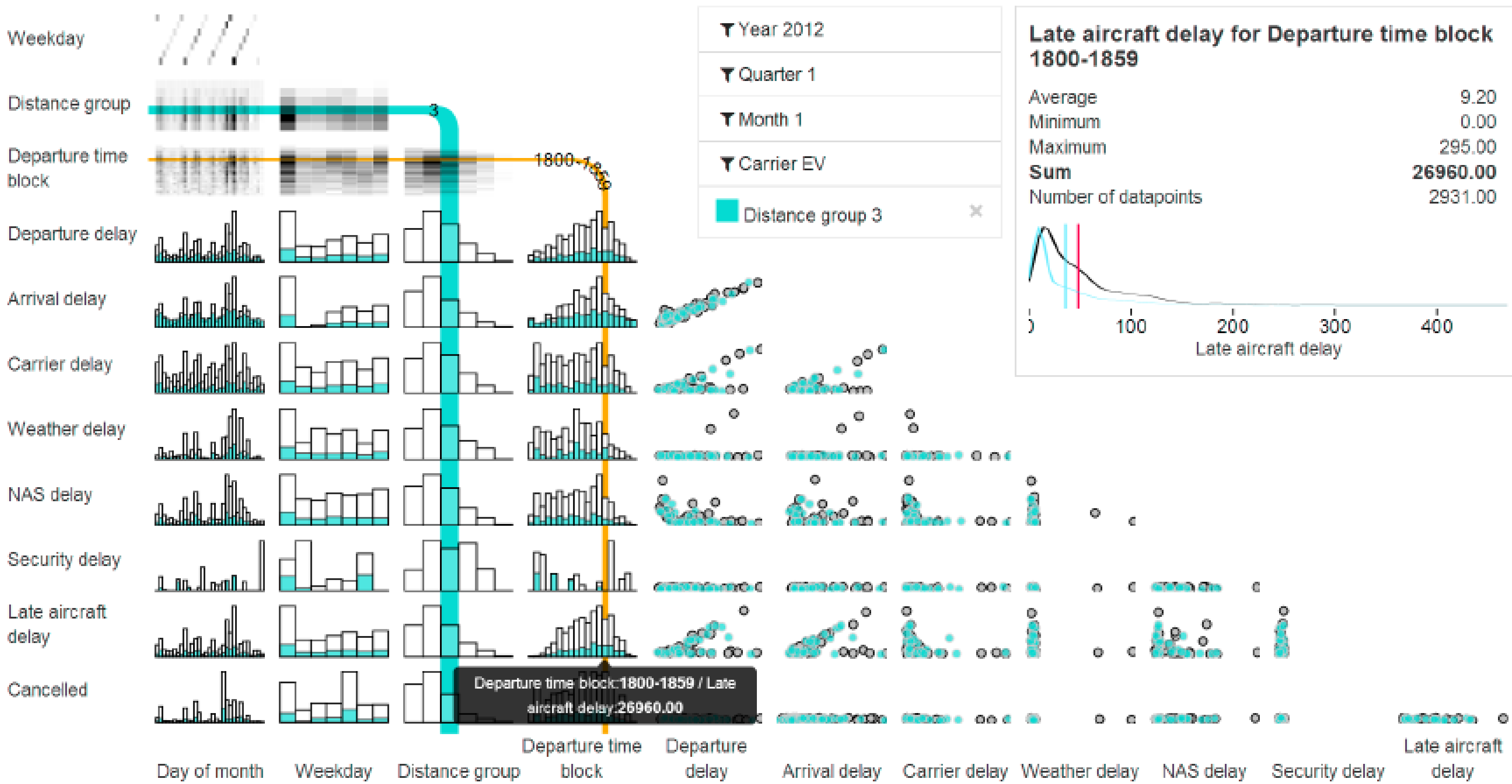
Scatterplots still show individual tuples (pairs of data points)

Images from GPLOM paper – Im, McGuffin, Leung[15]

On a larger scale

Images from GPLOM paper – Im, McGuffin, Leung[15]

GPLOM Tool allows for associative highlighting and filtering for additional exploration



The GPLOM tool was shown to reduce analysis time and was ranked as more fluid and easier to learn than dimensional stacking in Tableau by test subjects.

Animation[16]

- Study showed that animation of 2D dataset which added a time dimension through animation:
  - Was liked by users for data viewing data and helped with chunking, interpreting, expectations, comparisons, and focusing/filtering
  - However, not favored for grasping the whole or statistically analyzing the data values
  - All subjects said it helped them focus on changes in the data, and they used the viewer controls to changed the speed of the animation and to go back and forth and repeatedly view specific segments
  - Subjects wanted ability to bookmark interesting sections for review

Additional Reading

- Didn't have time to get into these, but also see
  - Article about making visualizations better with Gestalt Laws: http://sixrevisions.com/usability/data-visualization-gestalt-laws
  - The DataViz Catalogue: http://www.datavizcatalogue.com
  - Scagnostics – scatterplot clustering for high-dimensional data: http://www.cs.uic.edu/~tdang/file/ScagExplorer.pdf

References

1. Few, S. (2009). *Now you see it: Simple visualization techniques for quantitative analysis*. Oakland, CA.
2. Kennedy, J. (2012). Principles of Information Visualization Tutorial – Part 1 Design Principles. Retrieved April 20, 2015.
3. Green, M. (1998). Toward a Perceptual Science of Multidimensional Data Visualization: Bertin and Beyond. Retrieved April 20, 2015.
4. Few, S. (2012). *Show me the numbers: Designing tables and graphs to enlighten*. Burlingame, CA: Analytics Press.
5. Few, S. (2014, May 1). Why Do We Visualize Quantitative Data? Retrieved April 25, 2015.
6. Keim, D. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics, 8*(1).
7. Kosara, R. (2015, March 8). The Value of Illustrating Numbers. Retrieved April 12, 2015, from https://eagereyes.org/blog/2015/the-value-of-illustrating-numbers
8. Perry, C. (2013, October 12). What makes a data visualization memorable? Retrieved April 12, 2015, from http://www.seas.harvard.edu/news/2013/10/what-makes-data-visualization-memorable
9. Exploratory data analysis. (n.d.). Retrieved April 26, 2015, from http://en.wikipedia.org/wiki/Exploratory_data_analysis
10. Ros, I., & Hyland, A. (2013, April 9). When Creating Visualizations, Question Everything. Retrieved April 21, 2015, from https://hbr.org/2013/04/when-creating-visualizations-question-everything
11. Craft, B., & Cairns, P. (2005). Beyond Guidelines: What Can We Learn from the Visual Information Seeking Mantra? Retrieved April 19, 2015
12. Kosara, R. (2013, April 11). The Science of What We Do (and Don't) Know About Data Visualization. Retrieved April 21, 2015, from https://hbr.org/2013/04/the-science-of-what-we-do-and-dont-know-about-data-visualization/
13. Wildbur, P. (1989). *Information graphics: A survey of typographic, diagrammatic, and cartographic communication*. New York: Van Nostrand Reinhold.
14. De Oliveira, M., & Levkowitz, H. (2003). From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics, 9*(3), 378-394. Retrieved April 26, 2015.
15. Im, J., Mcguffin, M., & Leung, R. (2013). GPLOM: The Generalized Plot Matrix for Visualizing Multidimensional Multivariate Data. *IEEE Transactions on Visualization and Computer Graphics, 19*(12). Retrieved April 25, 2015.
16. Nakakoji, K., Takashima, A., & Yamamoto, Y. (2001). Cognitive Effects of Animated Visualization in Exploratory Visual Data Analysis. *Information Visualization*.

Questions?