SAS® GLOBAL FORUM 2015

The Journey Is Yours

Visualizing Your Big Data

What to Use, When & Why

Tricia Aanderud

zencos

SAS GLOBAL FORUM

3581-2015

What is Your Question My Dear?

Consider The Great Newmann – who has the same requirement as all big data analysts

I can’t answer until I know what the question is

Why do you want to analyze the data?

- **Exploration**
  Satisfy curiosity, gain understanding
- **Confirmation**
  Prove a theory
- **Improve**
  Be better than others!

The rules for analyzing big data are essentially the same as other datasets

Rule 1: Data Must Be Contextual

Consider Your Question

What influence did news events have on stock market trading in the 1940s

Consider Your Data

- All newspaper content since 2001
- Stock market transactions from 1990s
- Best selling candy in 1950s

Rule 3: Data must be Accurate

Consider: All these pregnant men!

“Even more striking, between 15,000 and 20,000 men have been admitted to *obstetric wards* each year since 2003...”

British Medical Journal Report
Review of 2009-10 hospital data

Who cares about your data if it doesn’t make sense or it is wrong?

Rule 2: Data must be Quantifiable

Determine if there is something to count in the data

Unstructured data has be to be categorized and counted. You have to have a form that you can visualize.

What do you know about the data?

707-29-0007	1961	12	Y	Y	0007		
302-38-0002	2010	350		Y	0075	SIRI	X
300-33-2292	1972	8	Y		0003	APPLE	
707-22-0007	1960	224		Y	0050	SIRI	JFK
202-38-0002			N		0010	IBM	LAX
300-33-2999	1975	300		N	0032	SIRI	ATL
505			Y	Y	0000	SIR	
202		250	N	Y	5255		
300		255	Y	N	4363	SIR	
Imagine there are eleventy-billion more rows from here							

Companies? Airports? Why in same file?

EPIC FAIL
UNKNOWN DATA

Tax ID Number? Machine code? Year born? Why?

Yes and no? What was asked? Why?

Without metadata, this information is useless because you don’t know when it was it collected, by whom … or even why

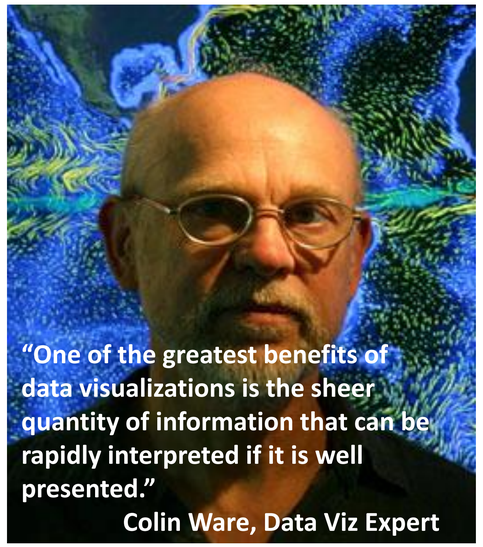


Image: http://ccom.unh.edu/

Visualization Types

- Line charts
- Bar charts
- Pie charts
- Spatial data
- Bubble plots
- Box plots
- Correlation Matrix
- Decision Maps
- Tag clouds

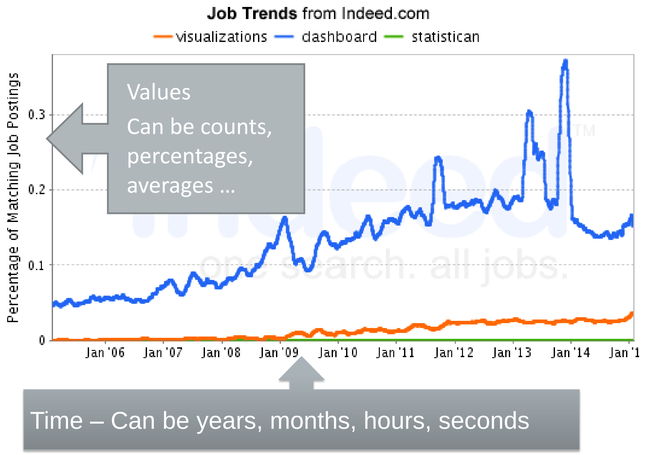
Same Rules Apply Across the Board

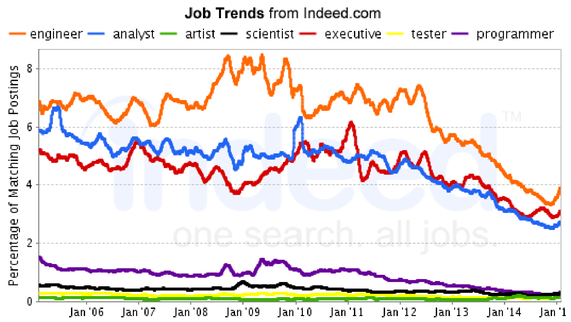
- Visualization should support your point – *make sure you know what it is*
- Keep it simple – less is more
- Consider your audience and what they understand
- Avoid colors, lines, patterns when it doesn't help

Line Charts Show Change Over Time

Guidelines

- Keep intervals even and in order
- Indicate missing values
- Use lines to connect points





Less is more …

Too many lines makes it harder to interpret the message –

Should I look for work as a scientist or not?

Area Charts are a form of line charts

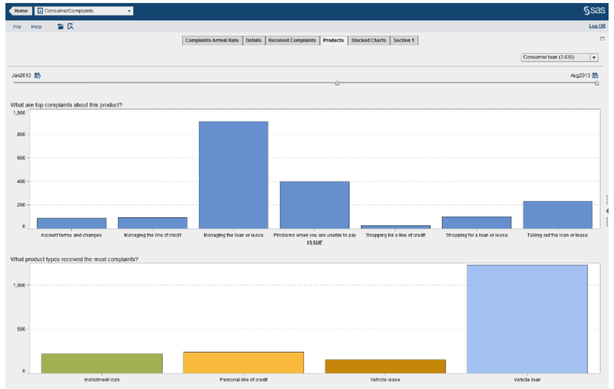
Did you notice the Y-Axis was different for each one?
Do you feel tricked or is it just showing how each company performed in same period?

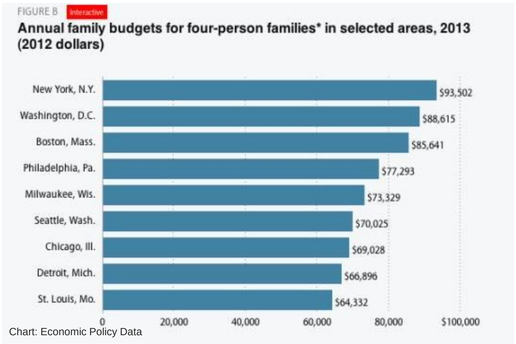
Bar Charts Compare Categorical Data

Vertical Bar Charts

- Values are on Y-Axis
- Shows more detail than line chart thus not intended for trend detection

- Easy to compare the categories
- Careful when there are more than 10 categories





Horizontal Bar Charts

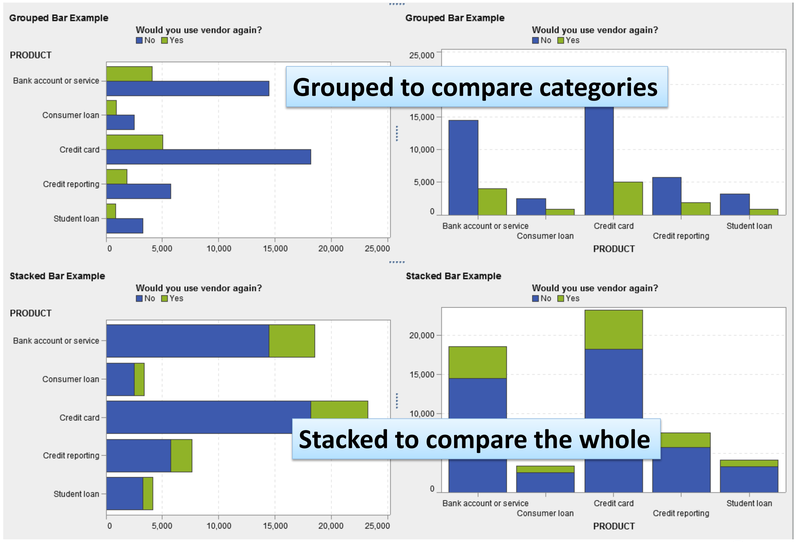
- Put values on X-Axis
- Allows you to rank values

- Use when you have really long labels
- More categories are easier to display with this chart type

Stacked vs Grouped

All of these charts use the same data – which one is easier to understand?

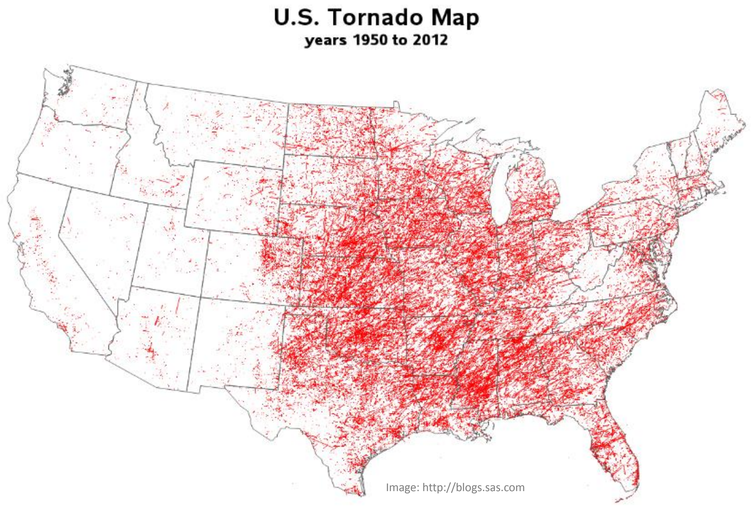
Maybe it's based on the question the user has?



"There are two goals when presenting data: convey your story and establish credibility."

-Edward Tufte
Data Visualization
Field Pioneer

Spatial Data Show Affected Geographical Area



Spatial Charts

- Showing data in a geographic area
- Usually a map but can be other areas
- Keep it simple by focusing on the area you want to show

Yikes! Tornado alley is no joke! But didn't take long to see it!

Tips

If the map supports a filter and interactive features – you can show more detail because users control what they want to see - Use SAS Visual Analytics!

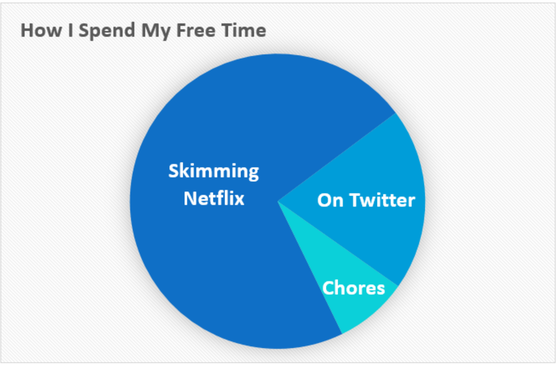
Pie Chart Show Parts as Whole

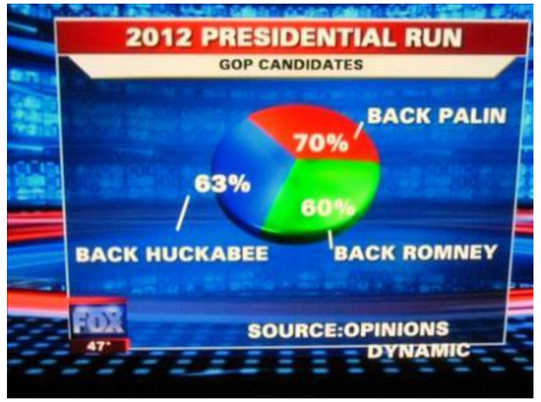
Guidelines

- Shows the parts as a whole
- Each slice is a category with its value
- Values expressed as percentage (usually)

Tips

- Limit to 3-4 categories
- Labels instead of legends
- Works best when one category is definitive





Slices should equal 100%

What is this chart trying to compare?

This figure totals 193% - so it doesn't make any sense. What exactly did Palin have the biggest share of?

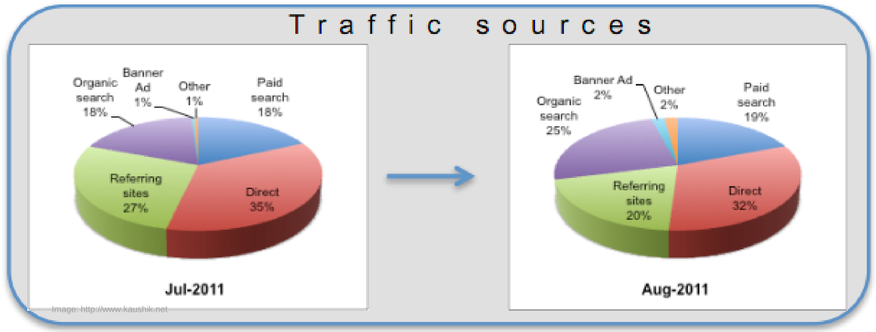
Visualization supports your point – but what was this author's point?

Avoid using pie charts for comparison

How long will it take you to determine the *second* largest source of traffic each month?

Wouldn't a bar chart work better?

Oh .. *and avoid 3D*



SAS® GLOBAL FORUM 2015

The Journey Is Yours

Tricia Aanderud
Zencos Consulting

Visit the Zencos blog for more data visualization tips

http://www.zencos.com/blog

- Connect to us on LinkedIn
- Follow us on Twitter

zencos

SAS GLOBAL FORUM