

Visualizations in the Analysis Workflow

SPE

Thomas Speidel, P.Stat., Data Scientist

2018/04/17

A copy of this presentation is available on GitHub: <https://github.com/tspeidel/SPE>

Views are my own.

Agenda

1. Visualizations: Objectives and Constraints
2. Visualizations for Data Storytelling
3. Univariate Visualizations
4. Bivariate Visualizations
5. Visualizing Models
6. Tips & Best Practices
7. What to Look for in Visualization Tools? An Opinionated List
8. References

Data Manifesto

Data, in and by themselves, do not directly create knowledge. While efforts to collect and store data have increased dramatically in recent years, nearly absent is a focus on knowledge creation. In the age of **Big Data**, the availability of vast amounts of information can coexist with the **absence of knowledge**.

Data cannot speak for themselves. It is in when we interpret data that knowledge is created. My focus is on **bridging the gap between data and knowledge creation**. That gap is filled by statistics and evidence based decision making.

1. Visualizations: Objectives and Constraints

Why Visualizations?

At their best, graphics are instruments for reasoning about quantitative information (Tufte, 1983)

Graphs are exceptionally powerful tools for data analysis (Cleveland, 1985)

There is no statistical tool that is as powerful as a well-chosen graph (Chambers, 1983)

Diagrams prove nothing, but bring outstanding features readily to the eye (Fisher, 1925)

Why Visualizations?

At their best, graphics are instruments for reasoning about quantitative information (Tufte, 1983)

Graphs are exceptionally powerful tools for data analysis (Cleveland, 1985)

There is no statistical tool that is as powerful as a well-chosen graph (Chambers, 1983)

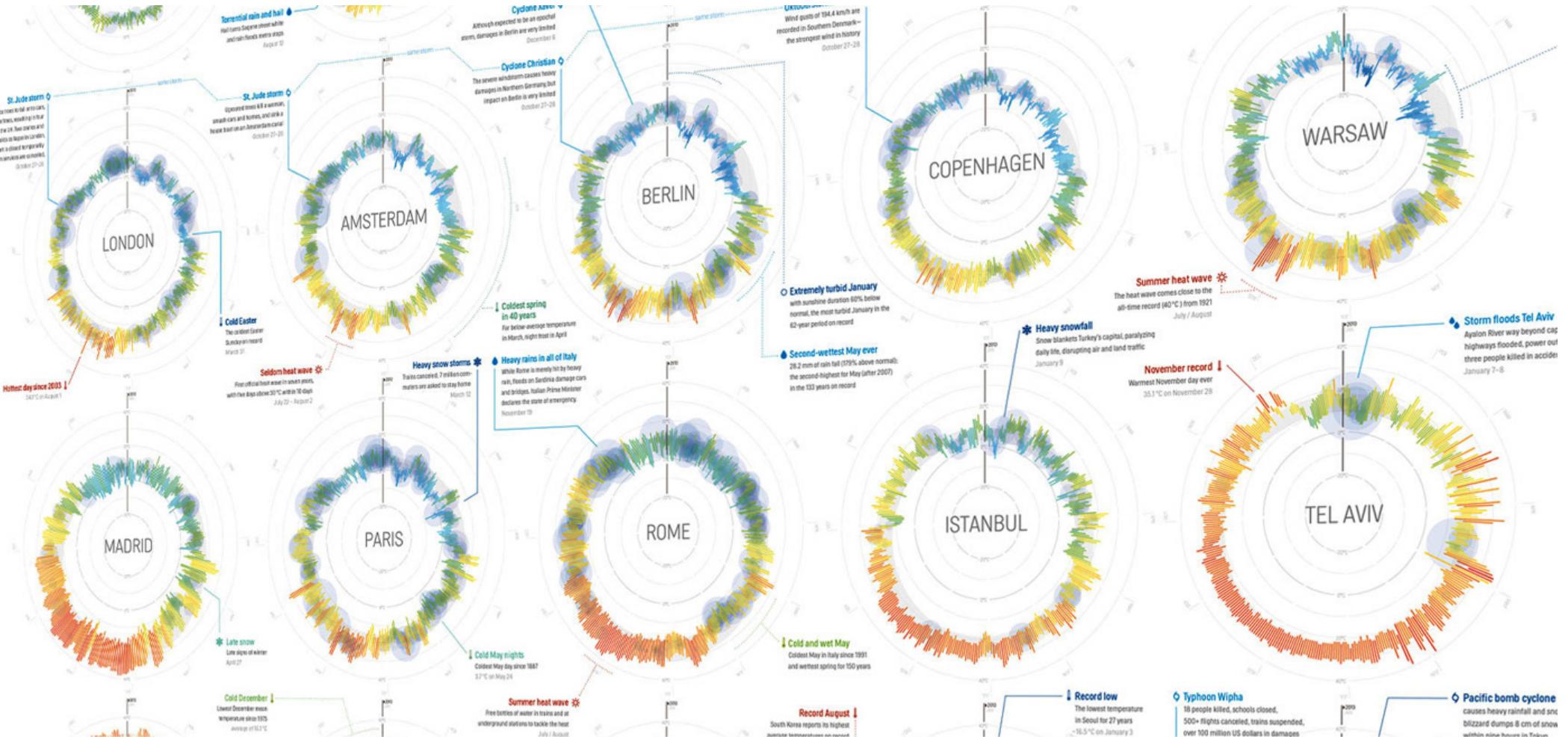
Diagrams prove nothing, but bring outstanding features readily to the eye (Fisher, 1925)

Two broad objectives:

- **Communication:** display a convincing pattern, attract reader's attention (first goal)
- **Discovery:** observe deviations from our expectations (second goal)

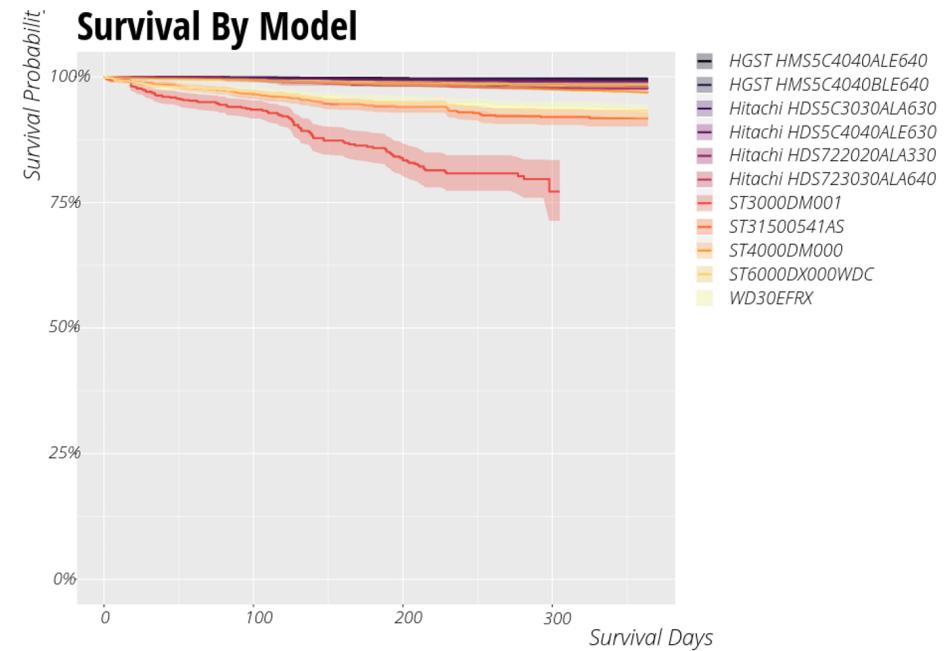
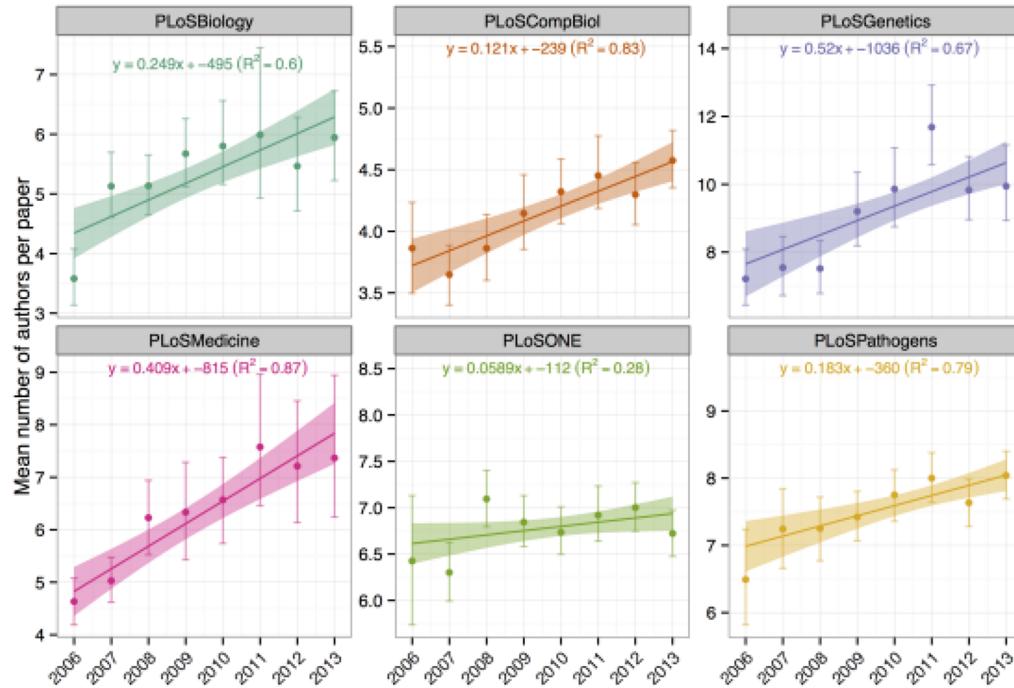
First Goal of Visualizations

To summarize, communicate and engage (telling a story)



Second Goal of Visualizations

To identify, understand, highlight features or relationships (helping the readers make up their mind)



Kaplan-Meier survival curves of time to failure for each hard drive model. Steeper curves indicate faster failure rates. By 300 days, only 77% of ST3000DM001 were still running, compared to >94% for all other drives. Could this be caused by the age of the drives?

To Persuade or to Inform?

We **cannot** achieve both effectiveliy at the same time!

To Persuade or to Inform?

We **cannot** achieve both effectiveliy at the same time!

To Persuade



To Persuade or to Inform?

We **cannot** achieve both effectiveliy at the same time!

To Persuade



To Inform

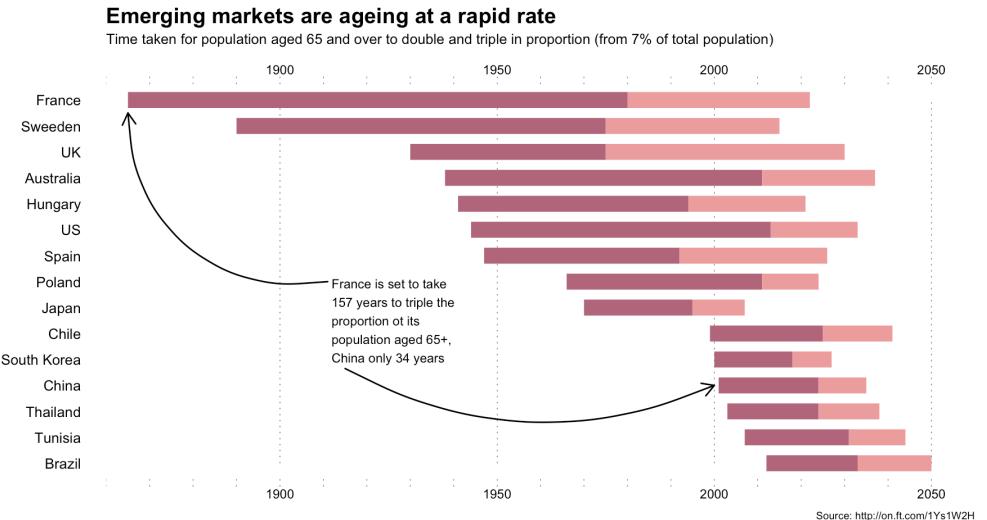
Nutrition Facts		
Serving Size 1 cup (228g)		
Servings Per Container 2		
Amount Per Serving		
Calories	250	Calories from Fat 110
% Daily Value*		
Total Fat 12g	18%	
Saturated Fat 3g	15%	
Trans Fat 3g		
Cholesterol 30mg	10%	
Sodium 470mg	20%	
Total Carbohydrate 31g	10%	
Dietary Fiber 0g	0%	
Sugars 5g		
Protein 5g		
Vitamin A	4%	
Vitamin C	2%	
Calcium	20%	
Iron	4%	
* Percent Daily Values are based on a 2,000 calorie diet. Your Daily Values may be higher or lower depending on your calorie needs.		
Calories	2,000	2,500
Total Fat	Less than 65g	80g
Sat Fat	Less than 20g	25g
Cholesterol	Less than 300mg	300mg
Sodium	Less than 2,400mg	2,400mg
Total Carbohydrate	30g	37.5g
Dietary Fiber	25g	30g

[1] Spiegelhalter, Pearson, Short. 2011 ↗

To Remember or to Understand?

Can we achieve memory **and** understanding at the same time?

1. **Titles and annotations** help convey the message of a visualization
2. If used appropriately, **pictograms** do not interfere with understanding and can improve recognition
3. **Redundancy** helps effectively communicate the message



[1] Borkin et al. 2015. *Beyond Memorability: Visualization Recognition and Recall* ↗

[2] Bob Rudis 2016. ↗

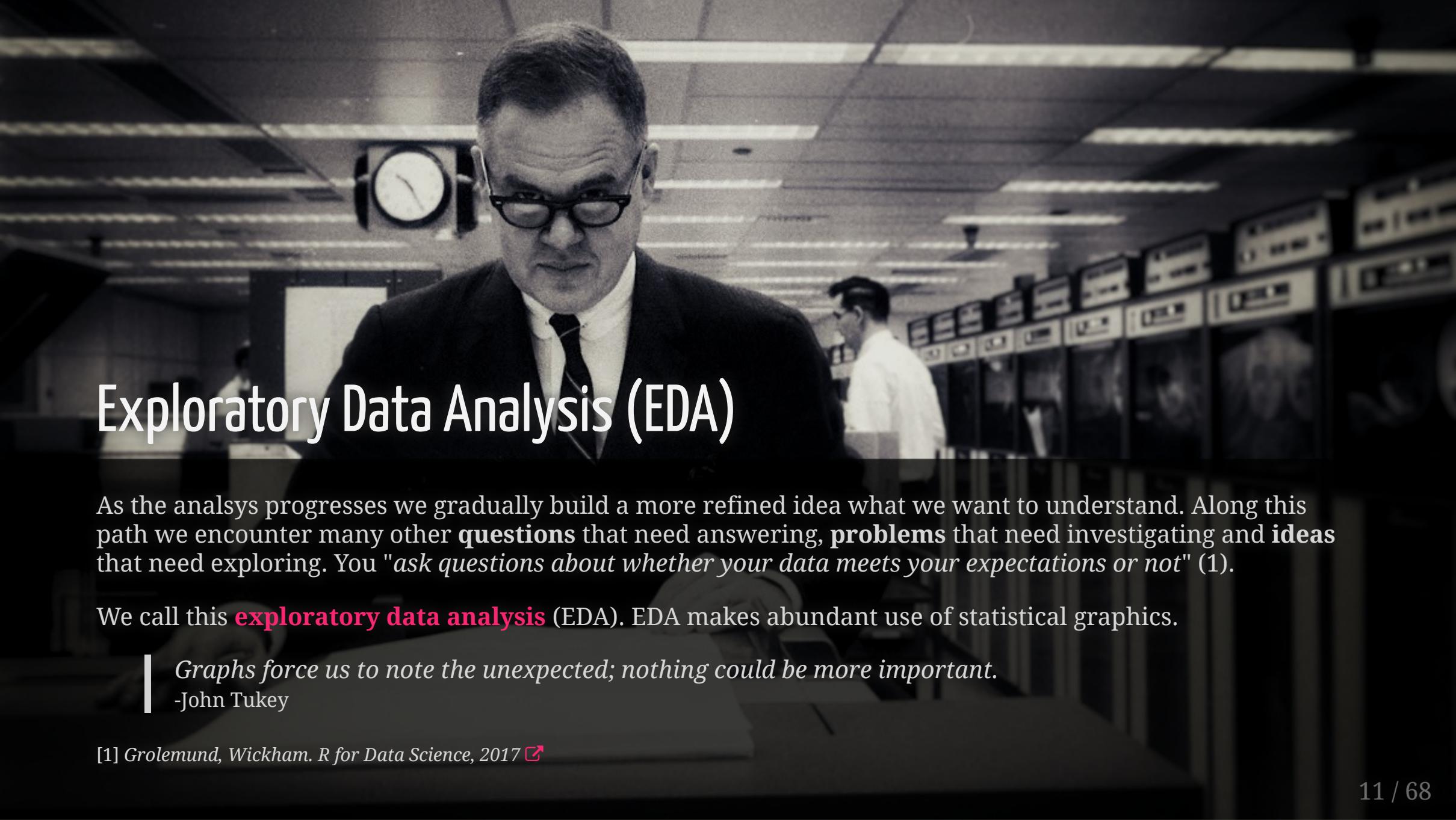
Example Dataset

The dataset originates from Doublet (2001).

well	depth	program	helium.porosity	gdens	kmax	k90	rtype	gr	rhob	pe	nphils	lld	lls
1509	6354.5	10-acre	8.6	2.84	6.57	1.52	1	26.81	2.720	3.411	0.0865	456.78	167.17
1509	6355.5	10-acre	5.2	2.85	0.09	0.07	1	22.83	2.745	3.326	0.0648	634.94	257.95
1509	7078.5	10-acre	1.7	2.83	0.40	0.12	1	21.09	2.758	3.146	0.0658	634.60	525.03
1509	7079.5	10-acre	1.9	2.82	0.25	0.17	1	24.99	2.761	3.237	0.0693	1003.90	811.04
1509	7080.5	10-acre	2.3	2.86	0.75	0.37	1	24.39	2.776	3.268	0.0648	1349.71	1204.42

[1] Thanks to D. Kaviani for pointing me to the data and for the preparatory work.

[2] Doublet, L.E. (2001) An Integrated Geologic and Engineering Characterization of the North Robertson (Clear Fork) Unit, Gaines County, Texas. Petroleum Engineering PhD thesis, Texas A&M University.



Exploratory Data Analysis (EDA)

As the analysis progresses we gradually build a more refined idea what we want to understand. Along this path we encounter many other **questions** that need answering, **problems** that need investigating and **ideas** that need exploring. You "ask questions about whether your data meets your expectations or not" (1).

We call this **exploratory data analysis** (EDA). EDA makes abundant use of statistical graphics.

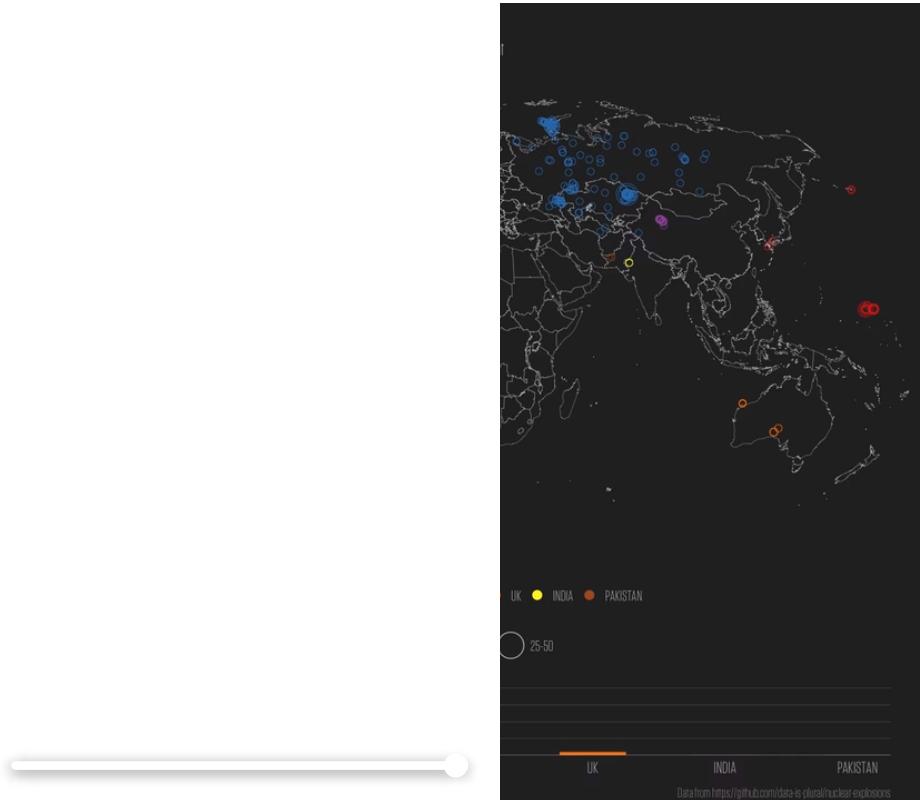
Graphs force us to note the unexpected; nothing could be more important.

-John Tukey

2. Visualizations for Data Storytelling

Telling a Story

Display a convincing pattern, attract reader's attention

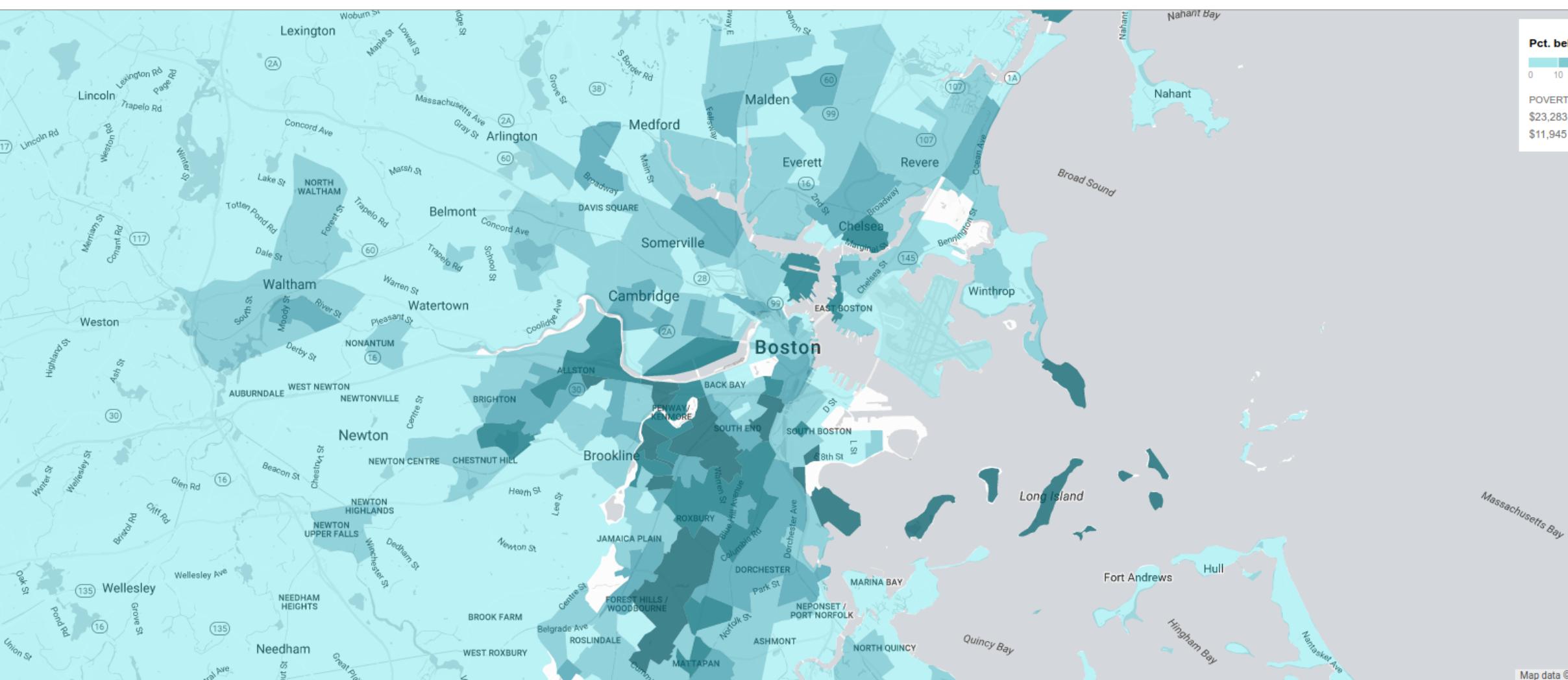


Mapping Poverty in America

Data from the Census Bureau show where the poor live.

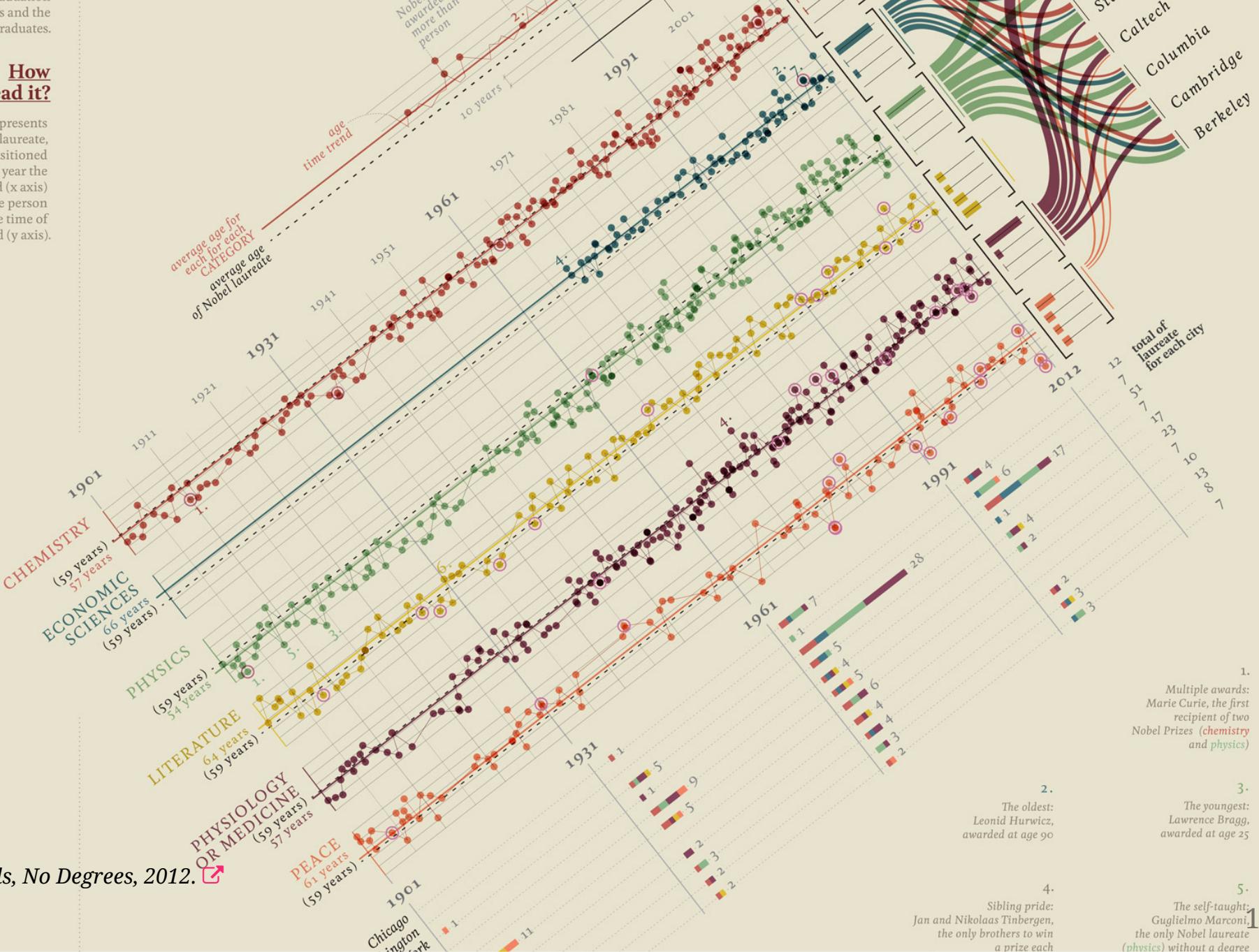
VIEW

Percentage below
the poverty line



How to read it?

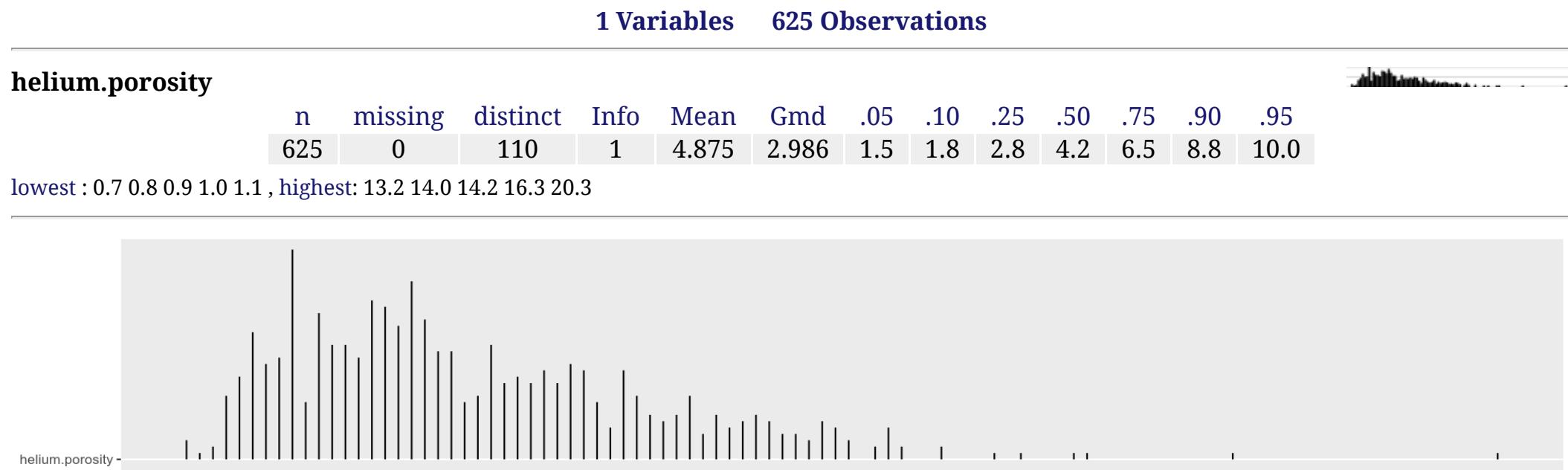
Each dot represents a Nobel laureate, each recipient is positioned according to the year the prize was awarded (x axis) and age of the person at the time of the award (y axis).



[1] Giorgia Lupi. Nobels, No Degrees, 2012. ↗

3. Univariate Visualizations

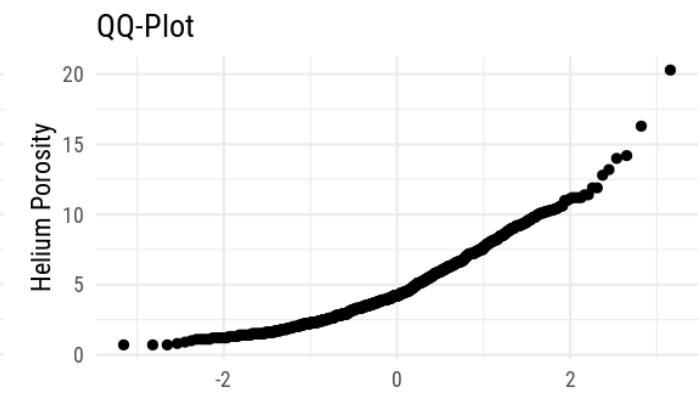
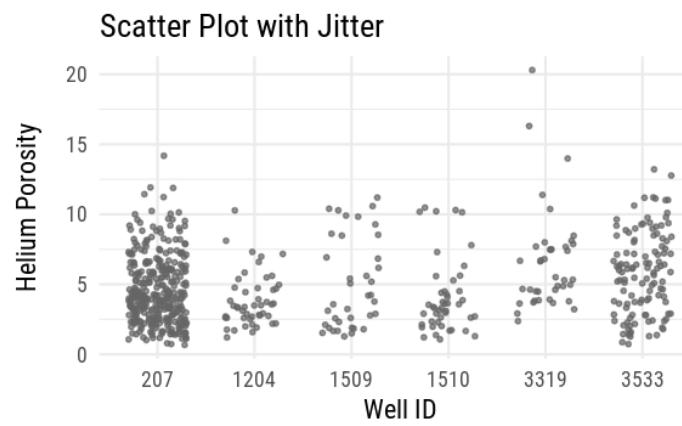
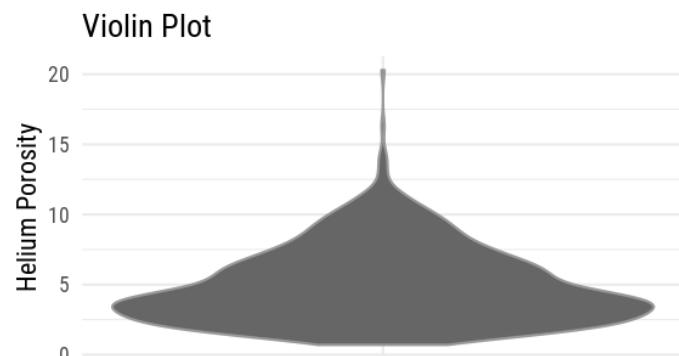
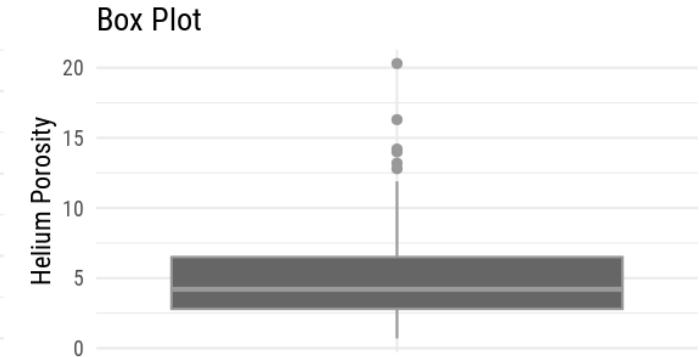
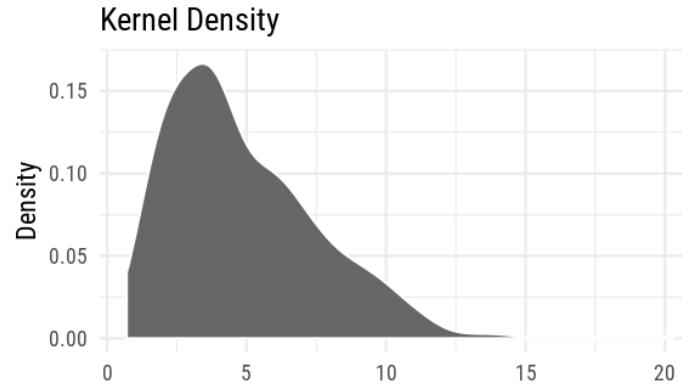
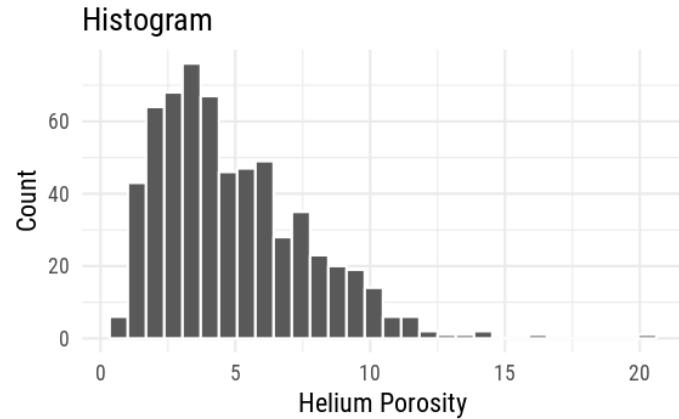
Summary Statistics



*Gmd stands for **Gini mean difference**, a scale independent measure of dispersion (it is the mean absolute difference between any pairs of observations).*

Info is a measure of how continuous the variable is.

Alternative Univariate Visualizations



4. Bivariate Visualizations

Numbers become evidence by being in relation to.

-Edward R. Tufte, Visual Explanations: Images & Quantities, Evidence & Narrative (1997)

Tables

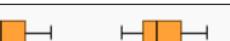
Baseline characteristics by rock type.

	1 N=80	2 N=180	3 N=37	4 N=26	5 N=42	6 N=229	7 N=31	Combined N=625	Test Statistic
depth	7008 7063 7128 6945 ± 308	6812 7008 7153 6916 ± 277	6858 7016 7092 6939 ± 246	6949 7151 7157 7002 ± 274	7225 7233 7244 7234 ± 14	6532 6880 7098 6823 ± 316	6791 6958 7159 6871 ± 296	6800 7042 7152 6910 ± 303	$F_{6\,618}=29$, P<0.001
helium.porosity	2.3 3.5 5.2 4.3 ± 2.9	3.0 4.2 6.0 4.7 ± 2.6	3.1 4.5 6.1 4.7 ± 2.1	4.3 5.8 7.5 5.9 ± 1.9	6.6 8.2 9.4 8.2 ± 2.7	2.5 3.8 5.8 4.4 ± 2.6	4.0 6.1 7.0 5.7 ± 2.3	2.8 4.2 6.5 4.9 ± 2.7	$F_{6\,618}=15$, P<0.001
gdens	2.840 2.845 2.860 2.843 ± 0.036	2.830 2.840 2.860 2.826 ± 0.193	2.830 2.840 2.850 2.843 ± 0.019	2.822 2.840 2.850 2.837 ± 0.032	2.723 2.740 2.750 2.745 ± 0.038	2.830 2.850 2.860 2.846 ± 0.020	2.830 2.840 2.850 2.837 ± 0.036	2.830 2.840 2.850 2.832 ± 0.108	$F_{6\,618}=21$, P<0.001
kmax	0.100 0.330 1.248 23.706 ± 199.208	0.070 0.250 1.088 1.983 ± 7.148	0.040 0.150 0.450 0.742 ± 1.465	0.092 0.210 0.697 1.236 ± 2.465	0.053 0.105 0.367 0.342 ± 0.532	0.090 0.280 1.120 2.018 ± 6.363	0.100 0.220 1.040 1.249 ± 2.397	0.070 0.240 1.020 4.525 ± 71.474	$F_{6\,618}=2$, P=0.063

a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables. $x \pm s$ represents $\bar{X} \pm 1$ SD.

Test used: Kruskal-Wallis test .

Combining Tables with Graphs

Summary Statistics and Box Plots for Temperature in Edmonton, Alberta from 1950s to 2000s									
Location	Month	Var	Decade	Mean	SD	Min	Max	Samples	Series
Edmonton, Alberta	Aug	Temperature	1950s	14.8	1.2	13.1	17		
Edmonton, Alberta	Aug	Temperature	1960s	15.7	1.61	13	18.2		
Edmonton, Alberta	Aug	Temperature	1970s	15.2	1.82	12.4	17.9		
Edmonton, Alberta	Aug	Temperature	1980s	15.5	2.03	13	18.8		
Edmonton, Alberta	Aug	Temperature	1990s	16	1.64	13.1	18.5		
Edmonton, Alberta	Aug	Temperature	2000s	15.3	1.22	14	17.3		

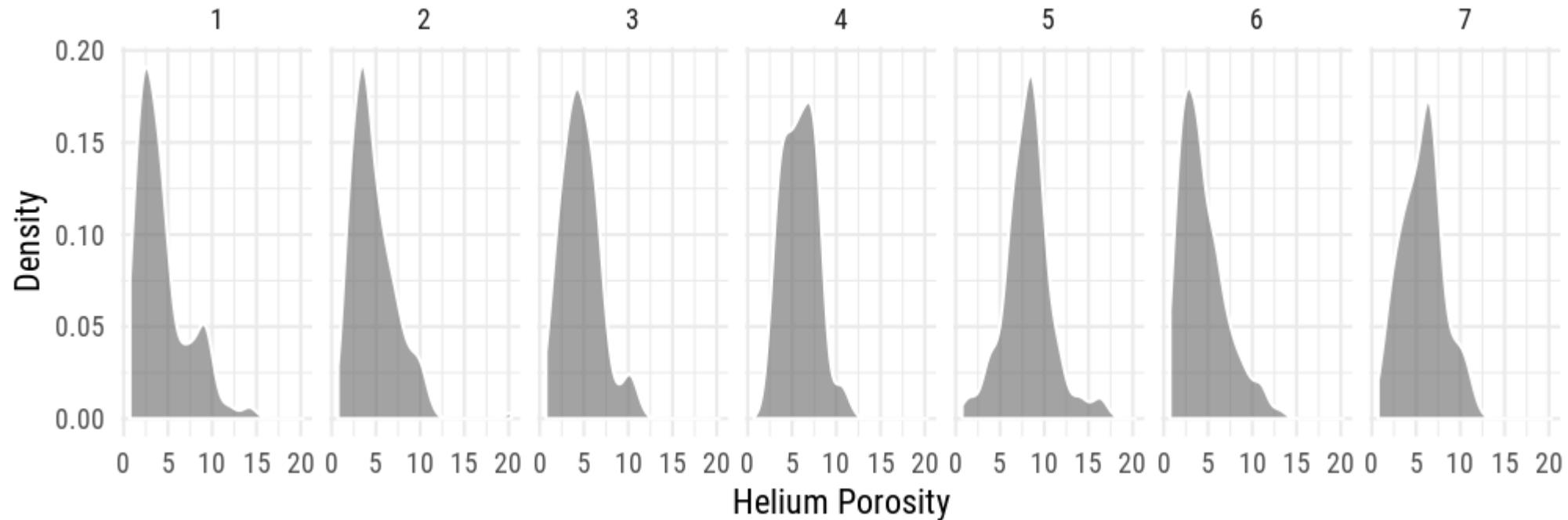
Show 10 entries Search:

Showing 1 to 6 of 6 entries Previous 1 Next

[1] Matthew Leonawicz ↗

Kernel Density

Kernel density plots have the unique advantage that they are self-standardizing. They are also more effective than histograms. This makes them useful to **compare distributions** when the units aren't the same (not here).



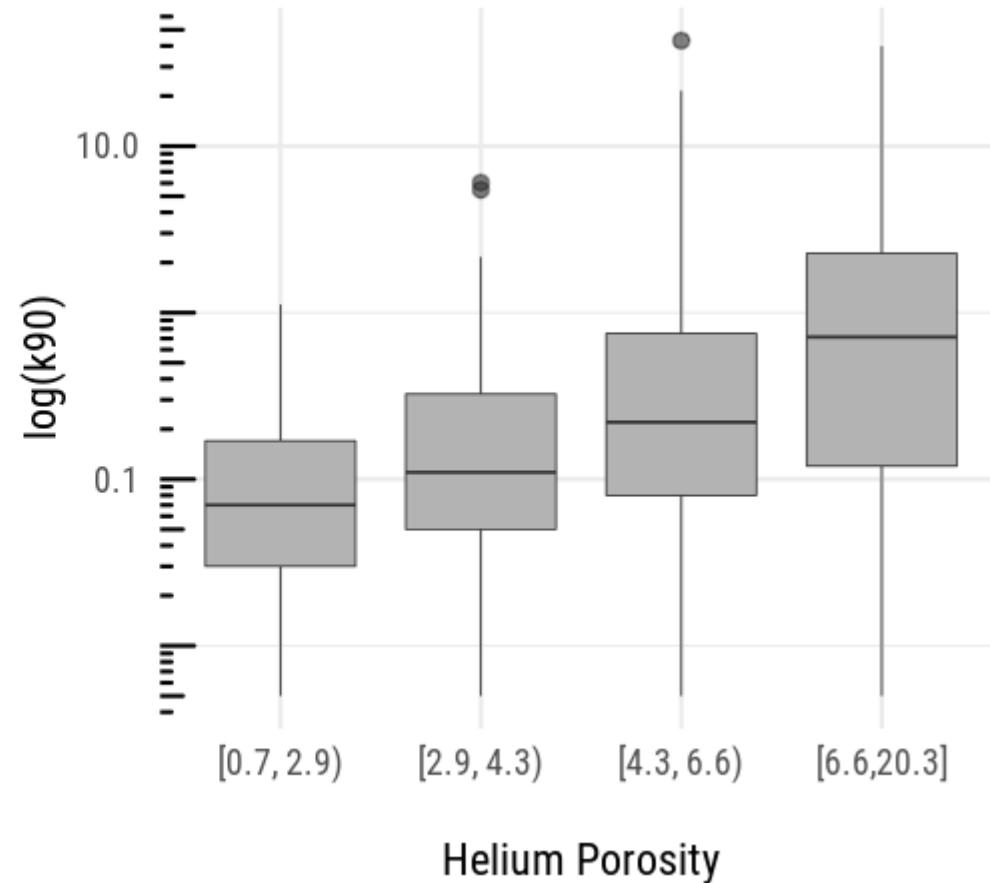
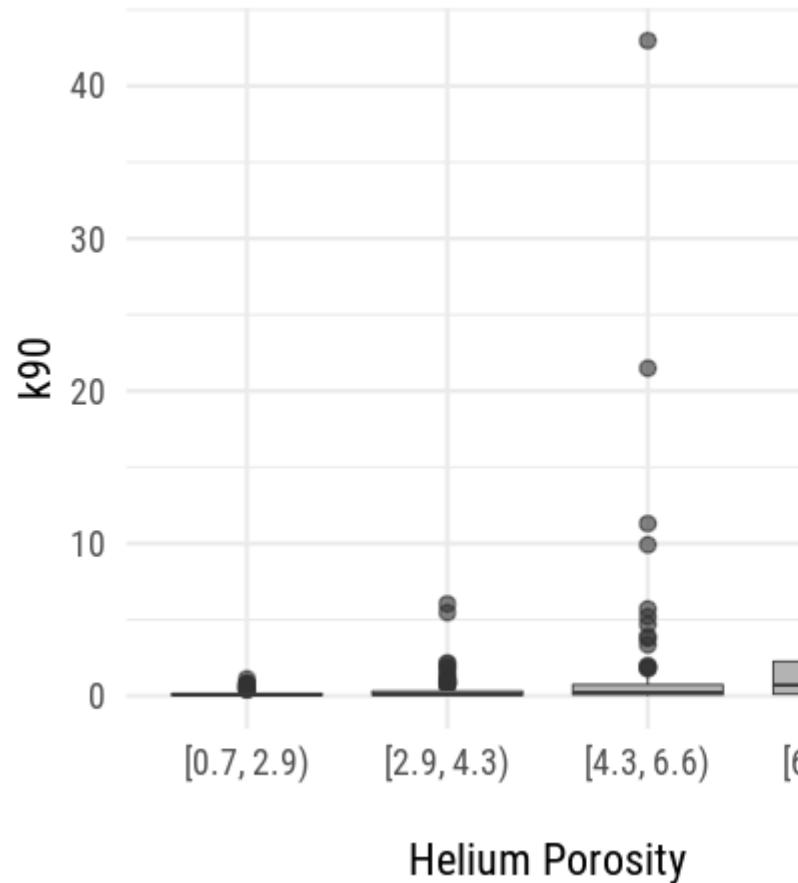
Scatterplot Matrix or Cross Plot

One of the very first plots you want to make. Shows all pairwise combinations. Here, I've cheated by removing a few outliers.

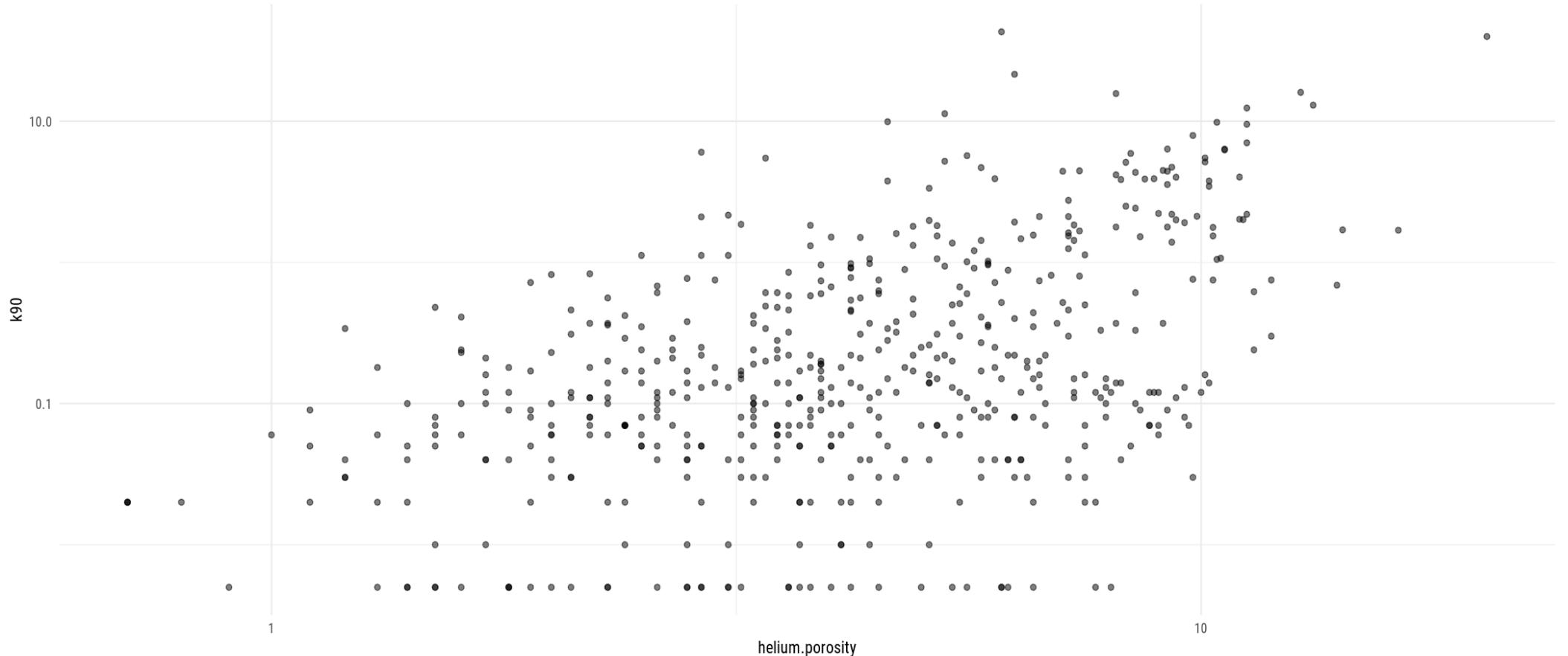


Boxplots

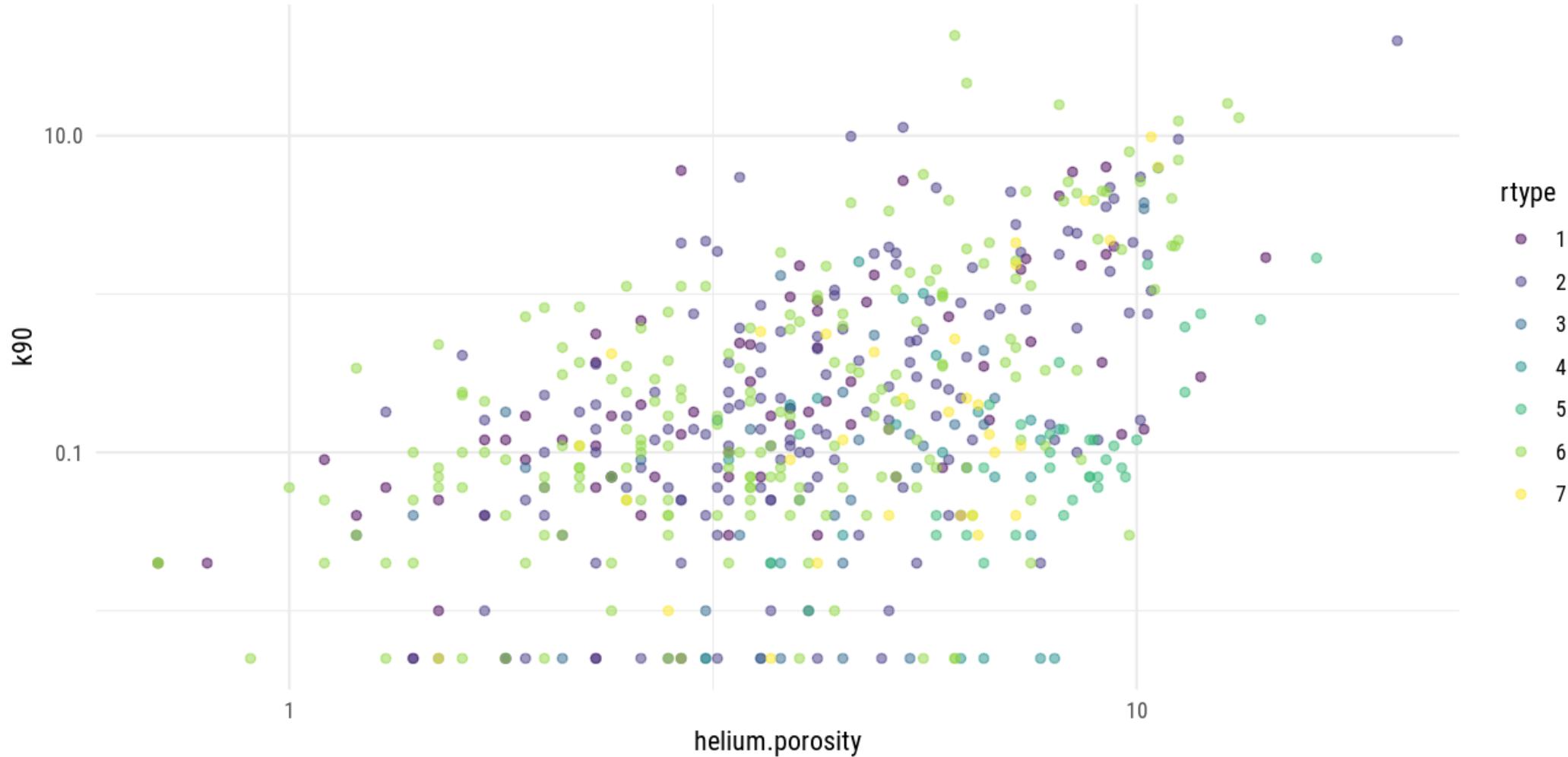
Here we *cut* (bin) helium.porosity into quartiles (P25, P50, P75).



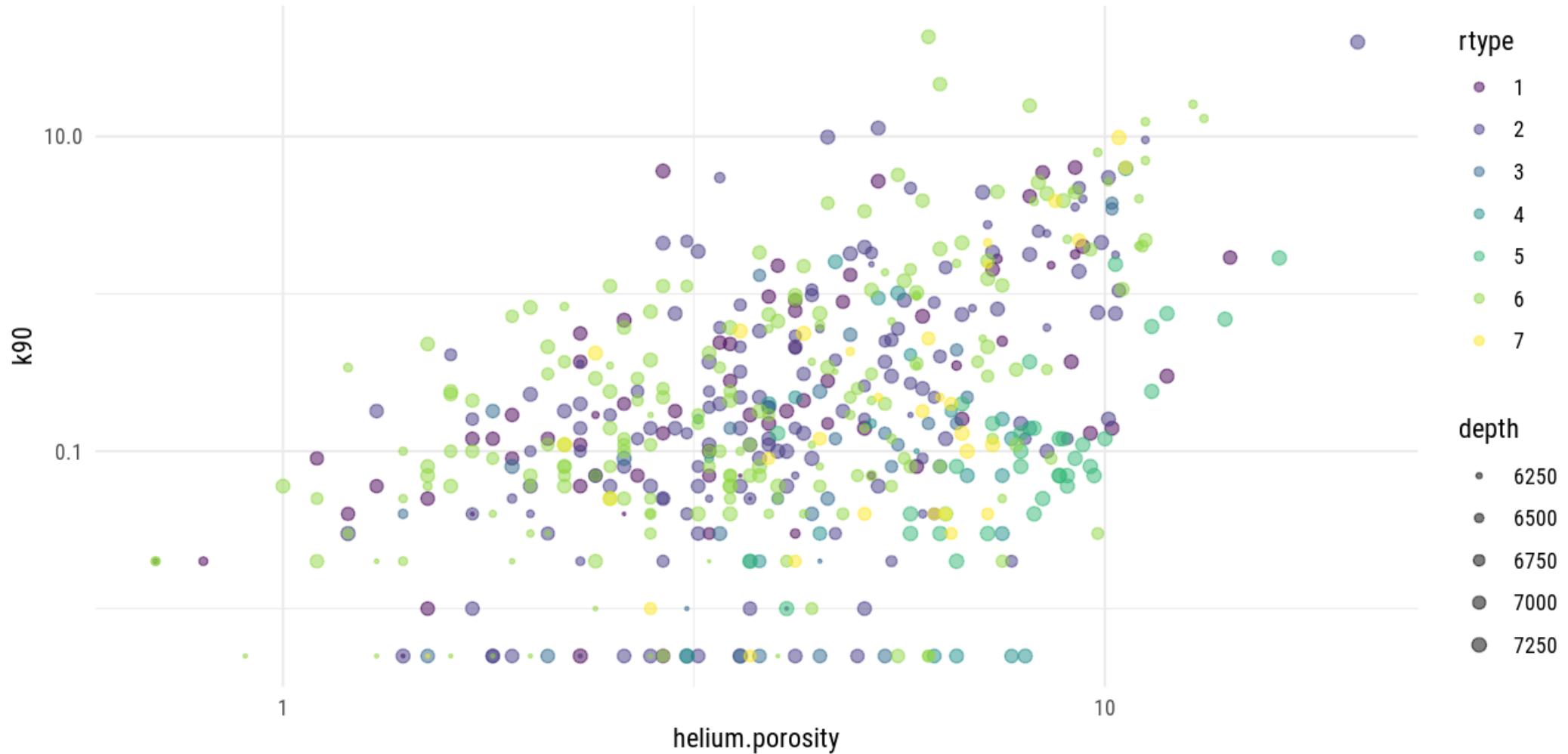
Scatterplots With 2 Variables



Scatterplots With 3 Variables



Scatterplots With 4 Variables



Correlation Matrix

A unitless measure of the strength of the association between pairs of variables.

		lls	lld	1				
			nphils	-0.7	-0.7			
			pe	-0.2	0.1	0.1		
			rholb	0.1	-0.7	0.7	0.7	
			gr	0.2	0	-0.2	0.2	0.2
			k90	-0.1	-0.1	0	0.3	0
			kmax	0.9	-0.1	0	0	0.3
			gdens	0	-0.1	0.2	0.4	0
			helium.porosity	-0.4	0.4	0.4	-0.2	-0.5
			depth	0.2	-0.4	0.1	0.1	-0.2
					-0.2	-0.4	0	0.1
						0.4	-0.3	-0.3
							0.4	-0.3

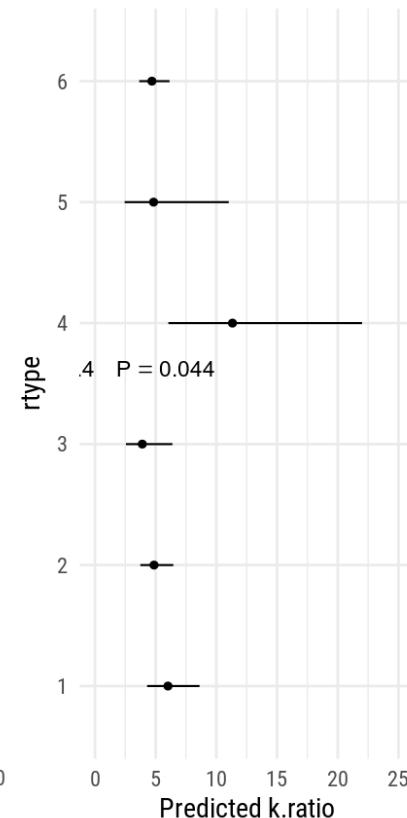
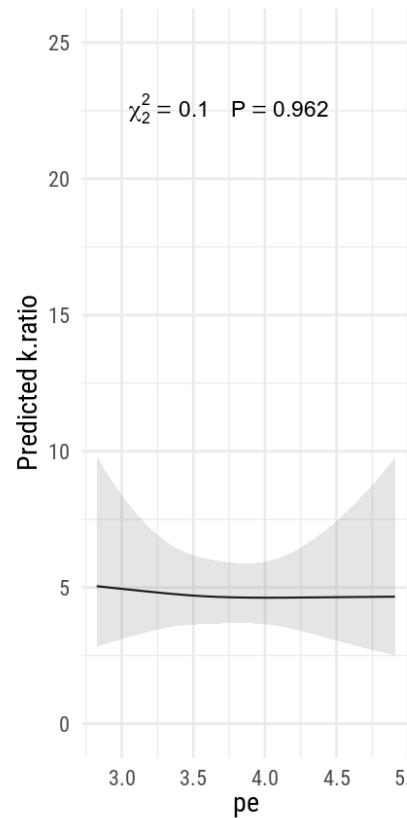
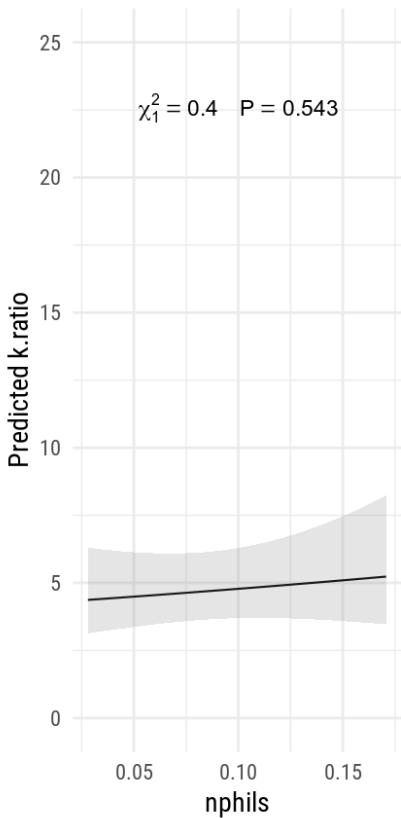
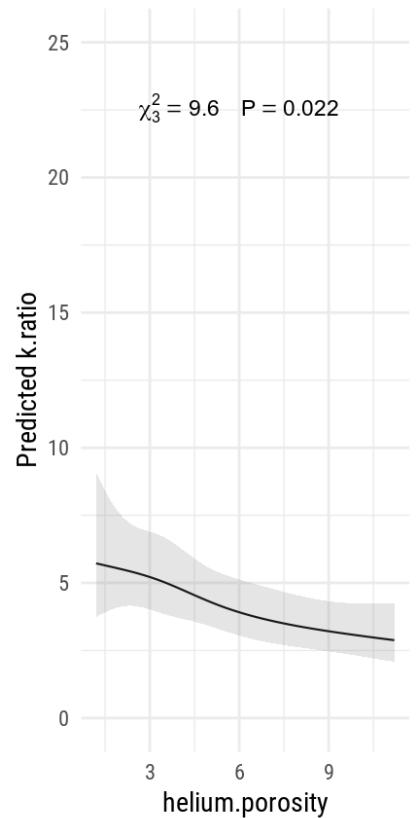
5. Visualizing Models

Graphs for Models

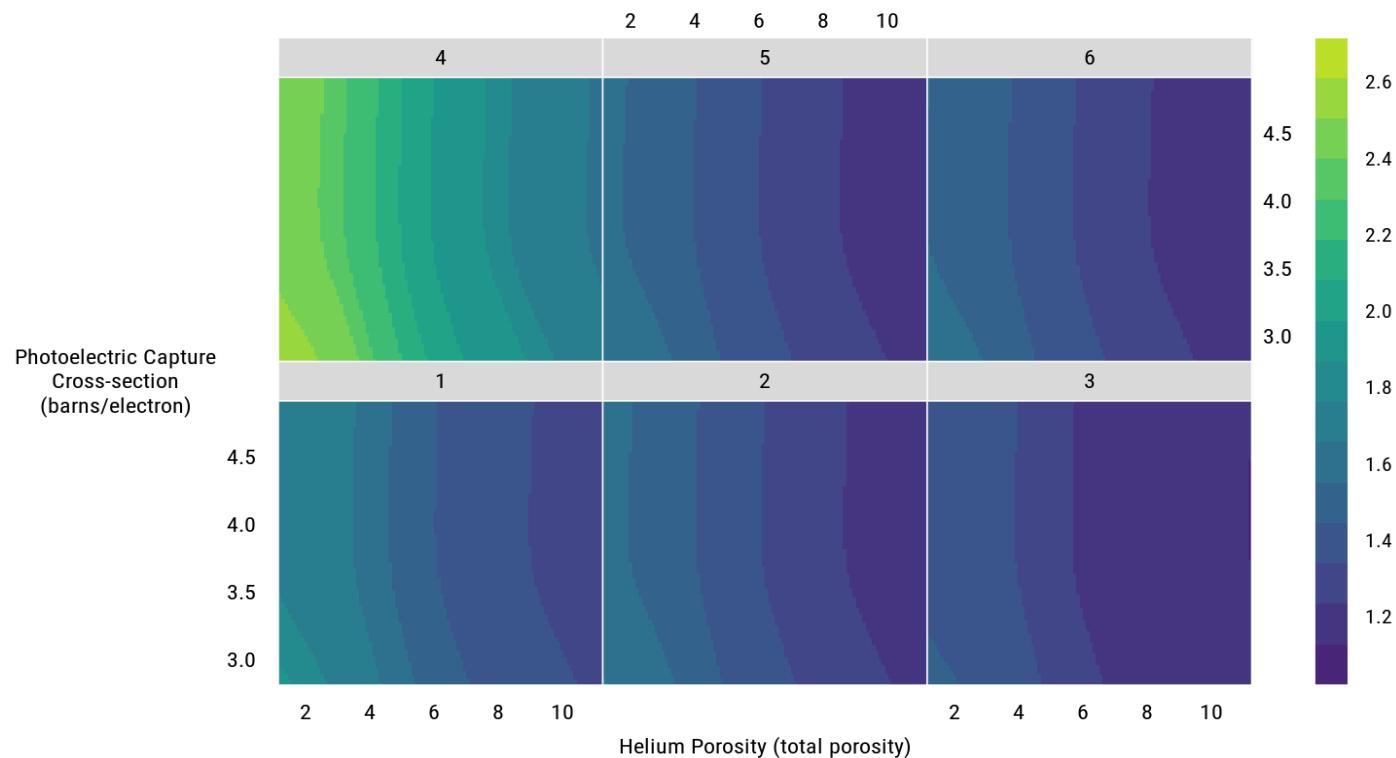
It's much easier to interpret a model by graphing it! Here's a regression model of `k.ratio` on several variables. Compare this:

Effects	Response: <code>k.ratio</code>						
	Low	High	Δ	Effect	S.E.	Lower 0.95	Upper 0.95
helium.porosity	2.80000	6.4000	3.60000	-0.47850	0.2176	-0.9049	-0.0521
<i>Odds Ratio</i>	2.80000	6.4000	3.60000	0.61970		0.4046	0.9492
pe	3.30600	3.8450	0.53980	-0.04675	0.1789	-0.3974	0.3039
<i>Odds Ratio</i>	3.30600	3.8450	0.53980	0.95430		0.6721	1.3550
nphils	0.06198	0.1115	0.04952	0.08786	0.1443	-0.1950	0.3707
<i>Odds Ratio</i>	0.06198	0.1115	0.04952	1.09200		0.8228	1.4490
rtype --- 1:6	6.00000	1.0000		0.34150	0.2686	-0.1849	0.8679
<i>Odds Ratio</i>	6.00000	1.0000		1.40700		0.8312	2.3820
rtype --- 2:6	6.00000	2.0000		0.05197	0.2484	-0.4348	0.5388
<i>Odds Ratio</i>	6.00000	2.0000		1.05300		0.6474	1.7140
rtype --- 3:6	6.00000	3.0000		-0.26770	0.3832	-1.0190	0.4834
<i>Odds Ratio</i>	6.00000	3.0000		0.76520		0.3611	1.6220
rtype --- 4:6	6.00000	4.0000		1.15000	0.4418	0.2845	2.0160
<i>Odds Ratio</i>	6.00000	4.0000		3.16000		1.3290	7.5110
rtype --- 5:6	6.00000	5.0000		0.04104	0.4848	-0.9092	0.9913
<i>Odds Ratio</i>	6.00000	5.0000		1.04200		0.4028	2.6950

... with this:

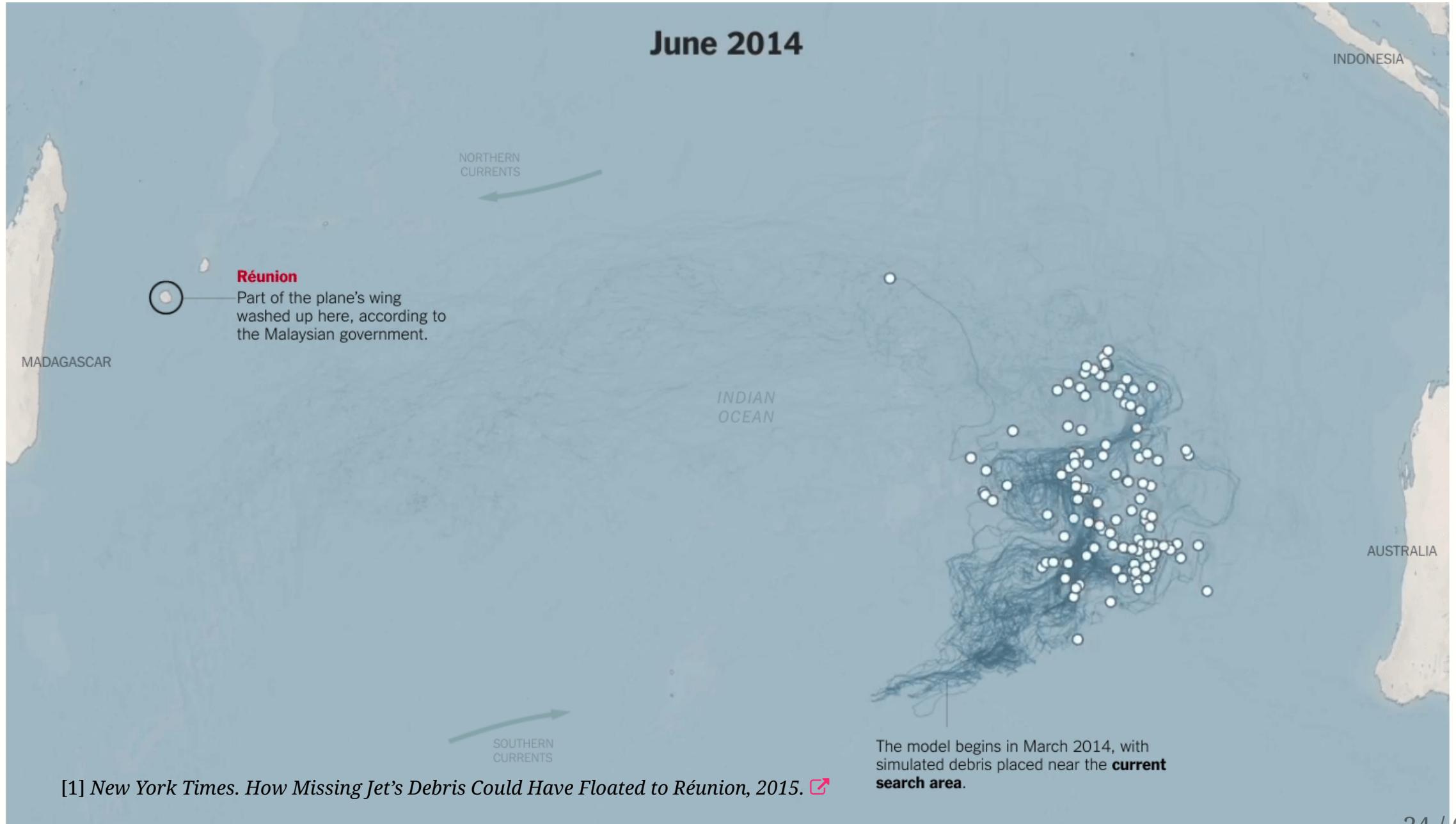


... or this:



Non-linear regression model showing the simultaneous effect, on the dependent variable permeability ratio (K_{max}/k_{90}), of changing two continuous variables (helium porosity and photoelectric capture cross-section log) while holding (marginalizing) all other variables fixed at their median.

June 2014



[1] New York Times. How Missing Jet's Debris Could Have Floated to Réunion, 2015. ↗

Visualizing Principal Components

Principal components allow us to describe our data with fewer new variables. These new variables are linear combinations of multiple variables. We call these new variables principal components.

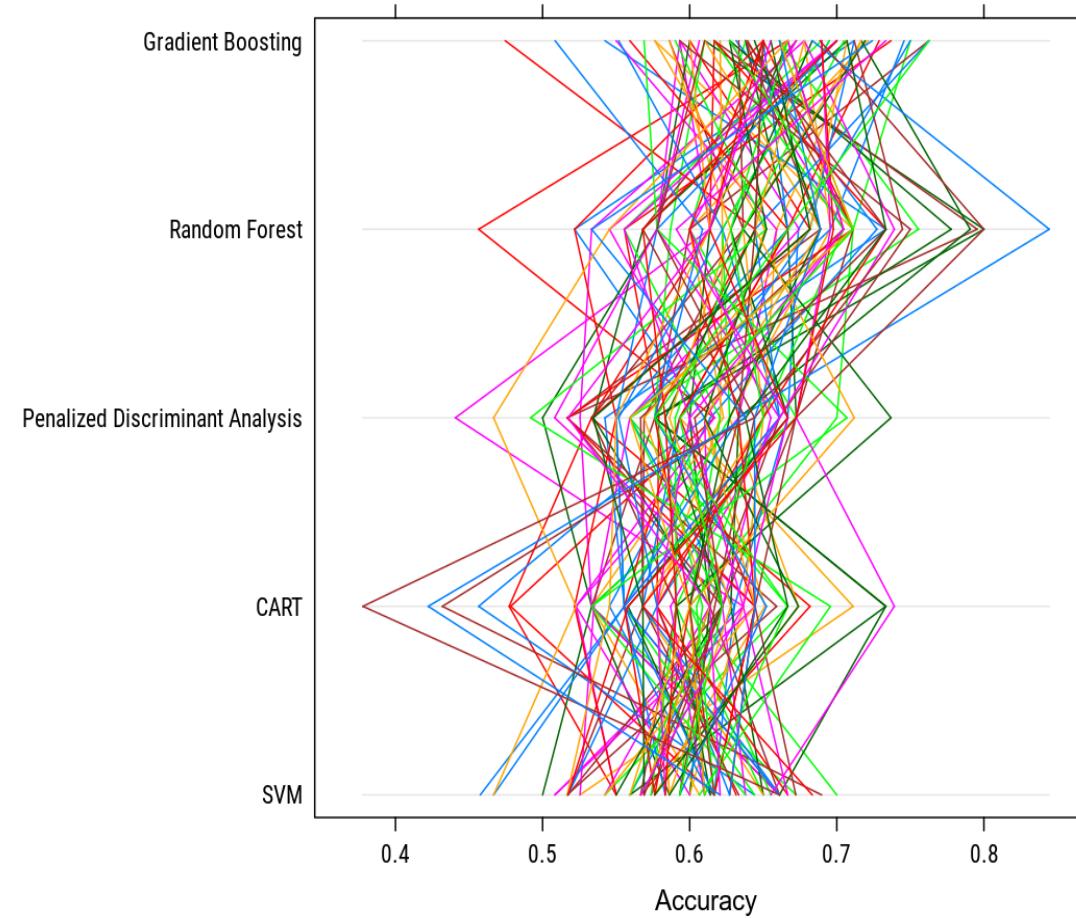
```
## Standard deviations (1, ..., p=6):  
## [1] 1.6000316 1.0858936 0.9579714 0.8818079 0.5800918 0.4784693  
##  
## Rotation (n x k) = (6 x 6):  
##          PC1       PC2       PC3       PC4       PC5  
## gr      0.22818282  0.137547398 -0.934774742 -0.2328681  0.03109767  
## rhob    0.43271157  0.453422040  0.265822874 -0.2945947  0.63176533  
## pe      0.03040547 -0.830000205  0.003589022 -0.4300128  0.34543234  
## nphils -0.48514647 -0.001751888 -0.221395134  0.5036248  0.67809619  
## lld     0.54616385 -0.153463229  0.042668603  0.3027975  0.12771134  
## lls     0.47550685 -0.251067310 -0.068447679  0.5733682 -0.06676349  
##          PC6  
## gr      0.003858898  
## rhob   -0.224930842  
## pe     -0.076991994  
## nphils 0.046493127  
## lld     0.753881245  
## lls     -0.610708536
```

Visualizing Principal Components

Visualizing CART Models

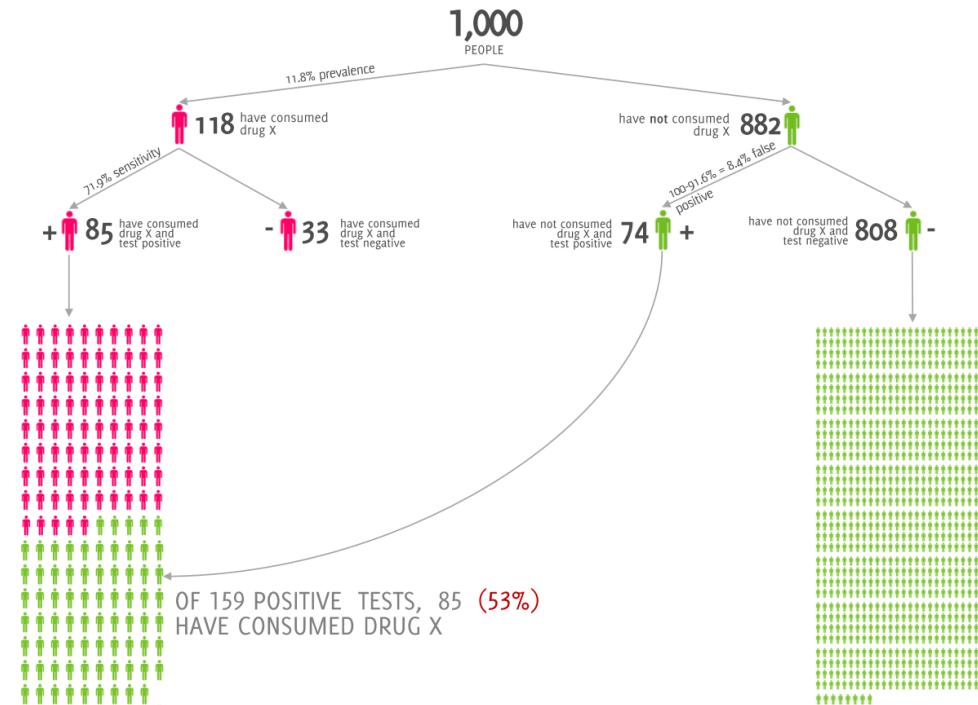
CART stands for **Classification and Regression Trees** a non-parametric type of model. Hypothetically, let's fit a model to predict rock type using depth and helium.porosity and visualize the results:

Visually Comparing Multiple Models



Diagrams to Explain

While not showing a model per-se, diagrams can help make unintuitive topics more transparent. Here, because of the low prevalence of drug consumption, we observe a high **false positive rate** (test says a person has consumed a drug, but in truth s/he has not), despite high test accuracy.



6. Tips & Best Practices

Information displays should be documentary, comparative, causal and explanatory, quantified, multivariate, exploratory, skeptical.

-Edward R. Tufte, Visual Explanations: Images & Quantities, Evidence & Narrative (1997)

General Graphing

— ANALYTICAL NOTES
— VISUALIZATION NOTES

By convention we put the variable of interest on the y-axis

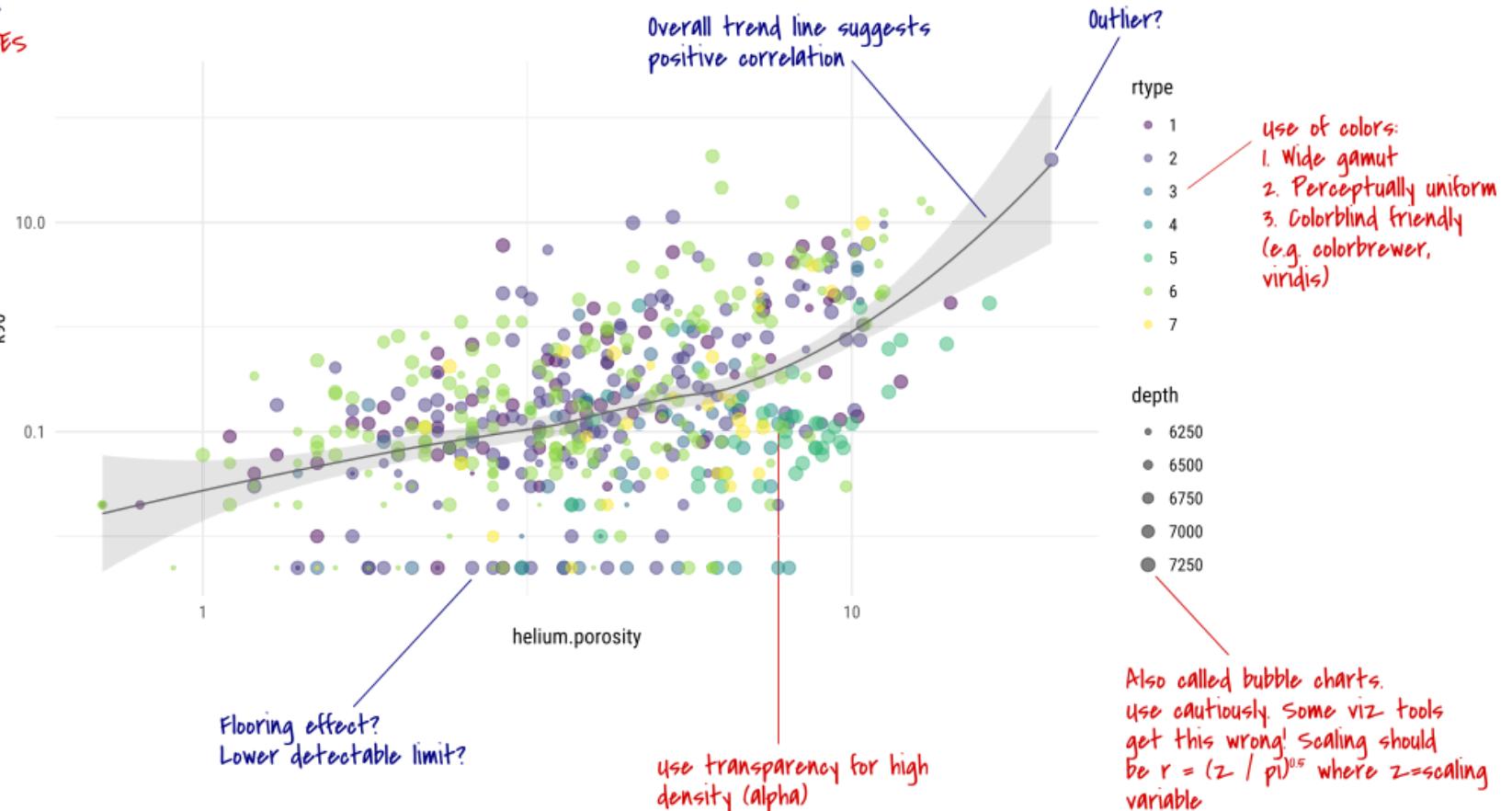


Table Tips

Sales in £ '000	QI	QII	QIII	QIV	Average
Sheffield	230	220	190	220	220
Leeds	280	190	220	340	270
Edinburgh	140	130	130	210	150
Hull	70	81	71	84	76
Swansea	62	66	62	77	67
Plymouth	41	44	33	50	42
Luton	23	27	23	27	25
Boston	31	29	25	29	29
Average	110	107	94	130	110

Decrease effective digits
(faced with long numbers
we all tend to be
non numerate)

Elements to be compared
across columns

Marginal averages gives
visual focus, provides
a summary and the
sorting order

Largest to smallest

White space aids
in readability

Do not sort
alphabetically

Right-aligned

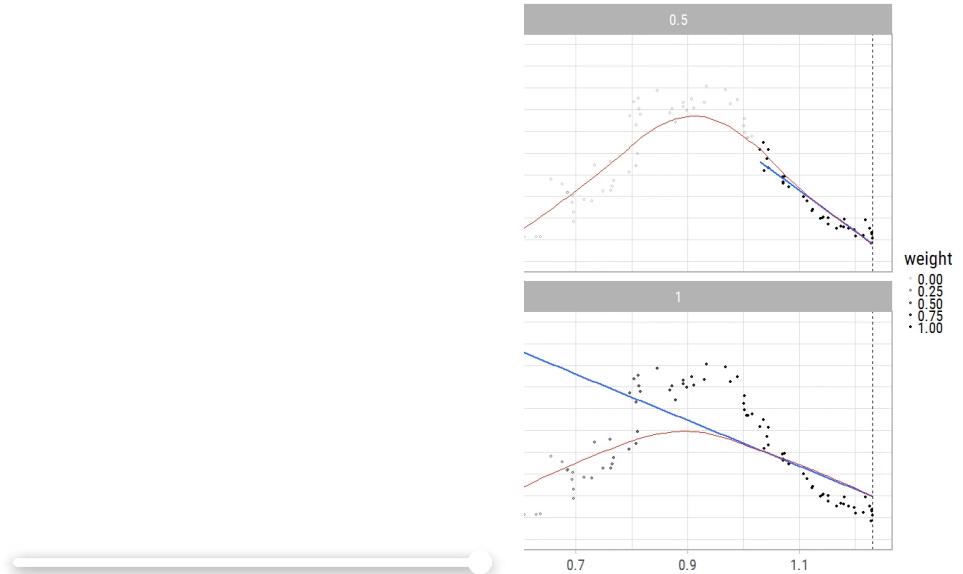
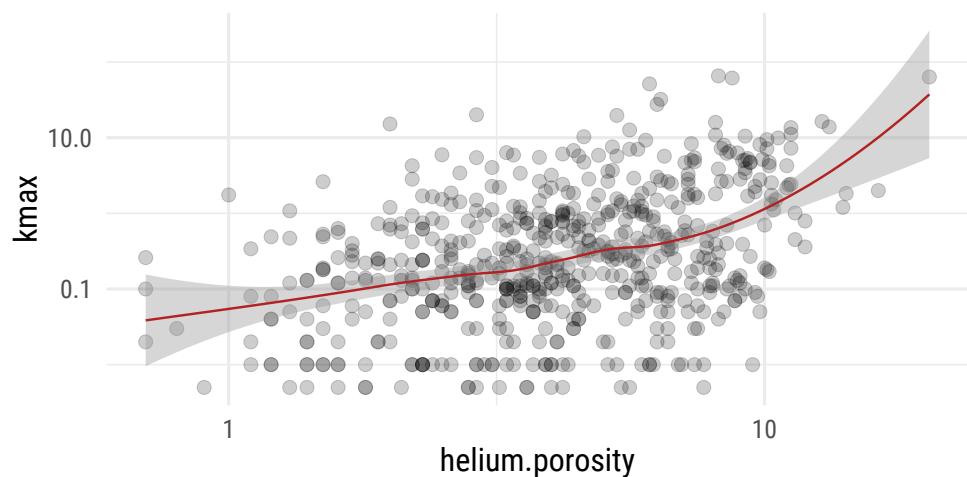
All numbers are rounded off
to a minimum (faced with long
numbers we all tend to be
non numerate)

[1] Recreated from Ehrenberg, 1981 ↗

Use (and abuse) Smoothers

Smoothers are incredibly useful in order to highlight possible **trends** in the data. One of the most popular and flexible smoother is Cleveland's **LOESS** (locally weighted least squares method).

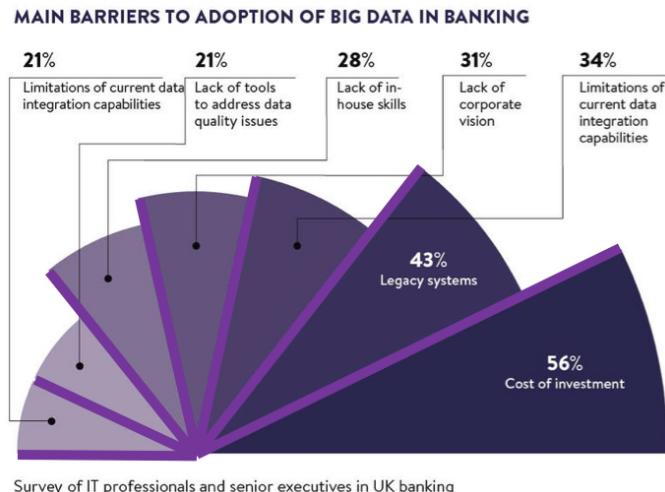
- Non-linear
- Built-in outlier control
- Can control degree of smoothness
- Works well even when one of the variables is binary



Data-to-Ink Ratio

Chartjunk

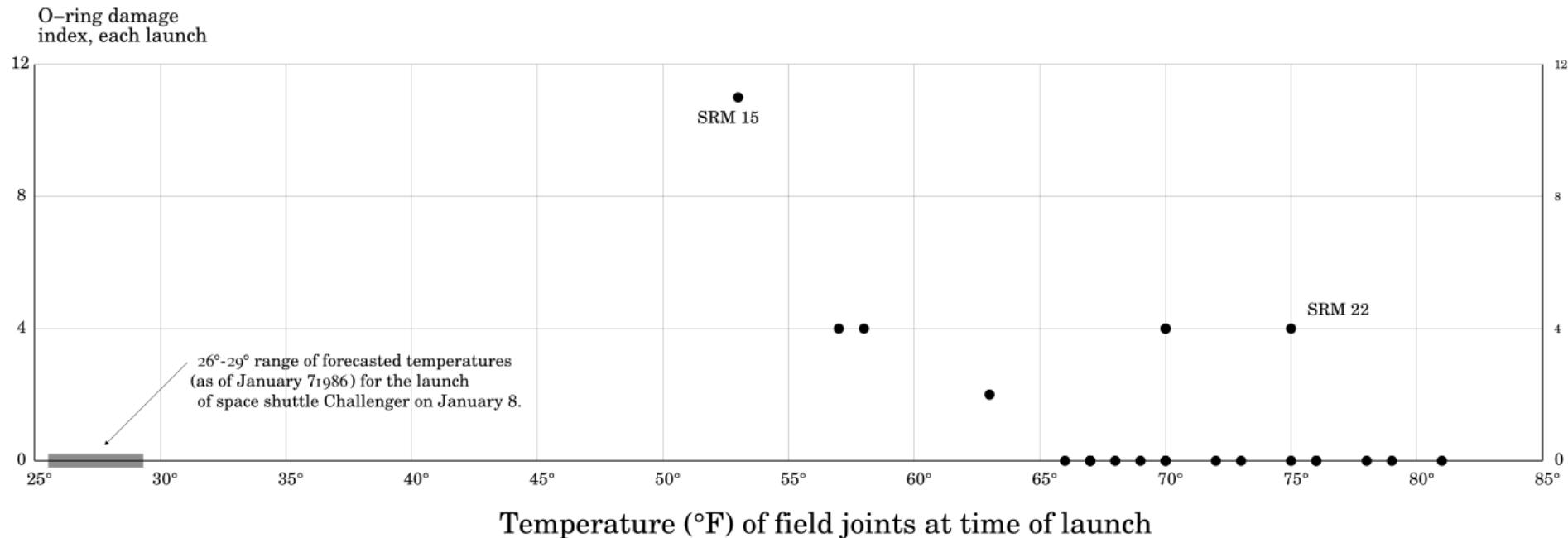
All visual elements in charts and graphs that are not necessary to comprehend the information represented on the graph, or that distract the viewer from this information. Instead, we want to achieve **rapid visual perception**. (Tufte, 1983, 2006)



Data-to-ink ratio

Of all the ink used in a graph, the amount of ink that **transports information**. We want this to be high. (Tufte, 1983, 2006)

Data-to-ink ratio

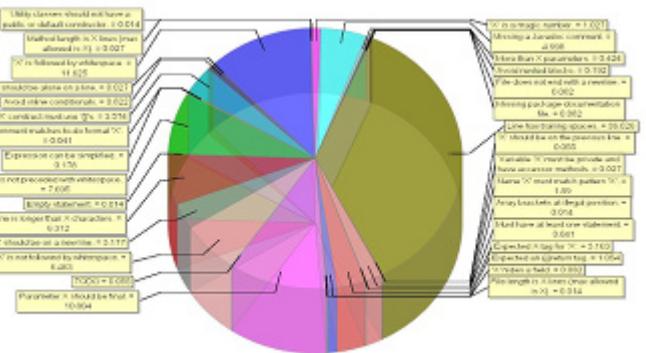


[1] Recreated from Tufte, 1983. Visual Explanations: Images and Quantities, Evidence and Narrative ↗

Avoid Pies

Why

Plenty of scientific evidence shows that humans are **bad at judging size and making comparisons from angles**. Pie charts do not facilitate comparison when slices are close in size nor can one sort numbers.



Exceptions

Ok to use when there are few well distinct slices to represent. Because we are only good at judging slices that are $\frac{1}{2}$ or $\frac{1}{4}$ of the total, we should aim to have **2-4 slices** at most.

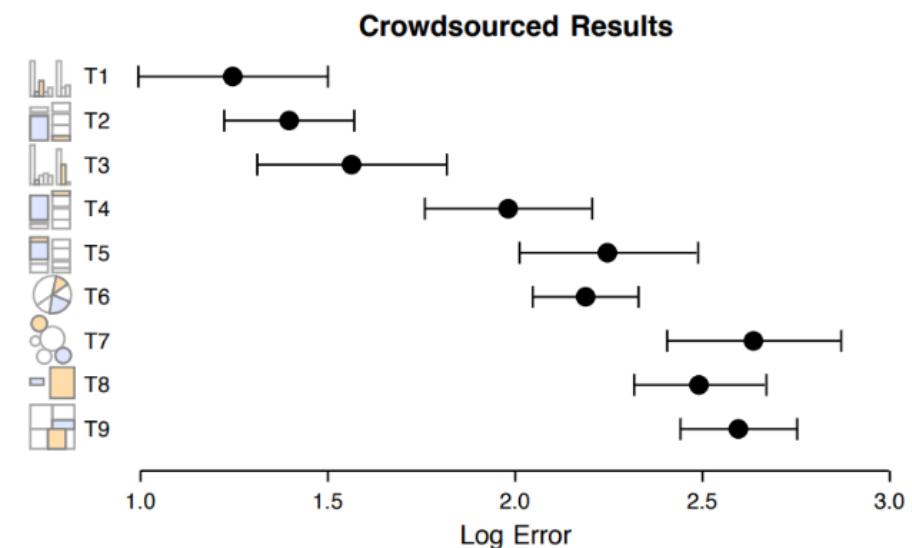
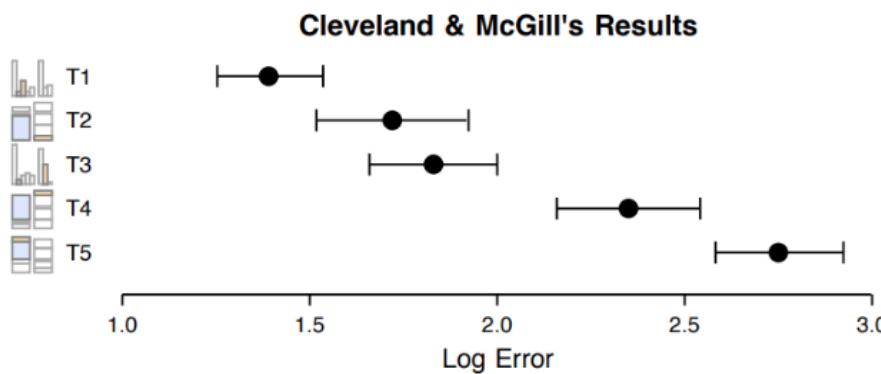
Avoid the **doughnut types**: they remove the angular information we need to draw visual comparisons.

Alternatives

Use bar charts or dot plots.

The Type of Chart Affects Visual Judgement

Experiments performed to quantify visual judgment of graphs: 25 years later, results have not changed.



[1] Cleveland and McGill, 1984 [↗](#) and Heer and Bostock, 2010 [↗](#)

Avoid 3D

Why

Removes clarity, adds obfuscation, inhibits comprehension, does not help retain information, third dimension is usually nonexistent (and if it did exist, you would try to avoid it).

Exceptions

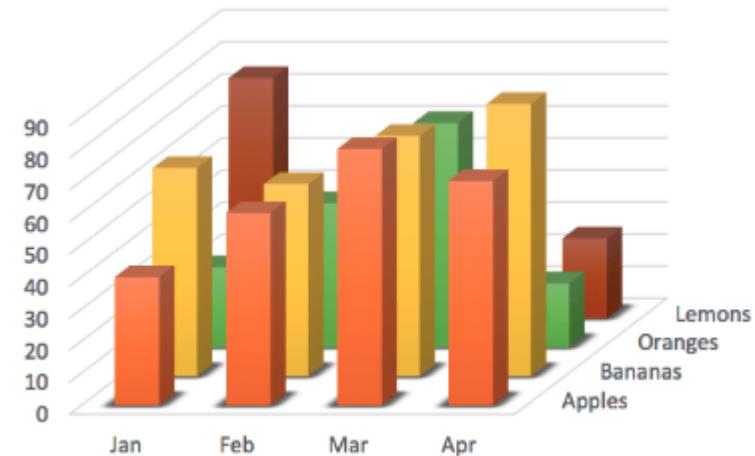
Spatial data, interaction surfaces.

Advantages

A reader is more likely to remember a 3D graph than a plain graph.

Disadvantage

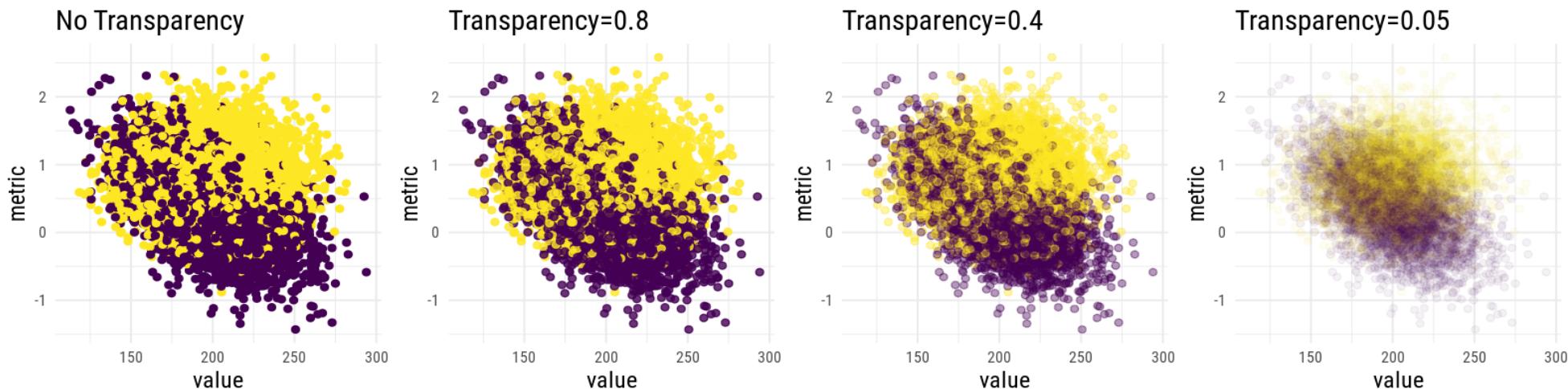
A reader is more likely to be confused by a 3D graph than a plain graph.



Visualizing High Density Data

Good...

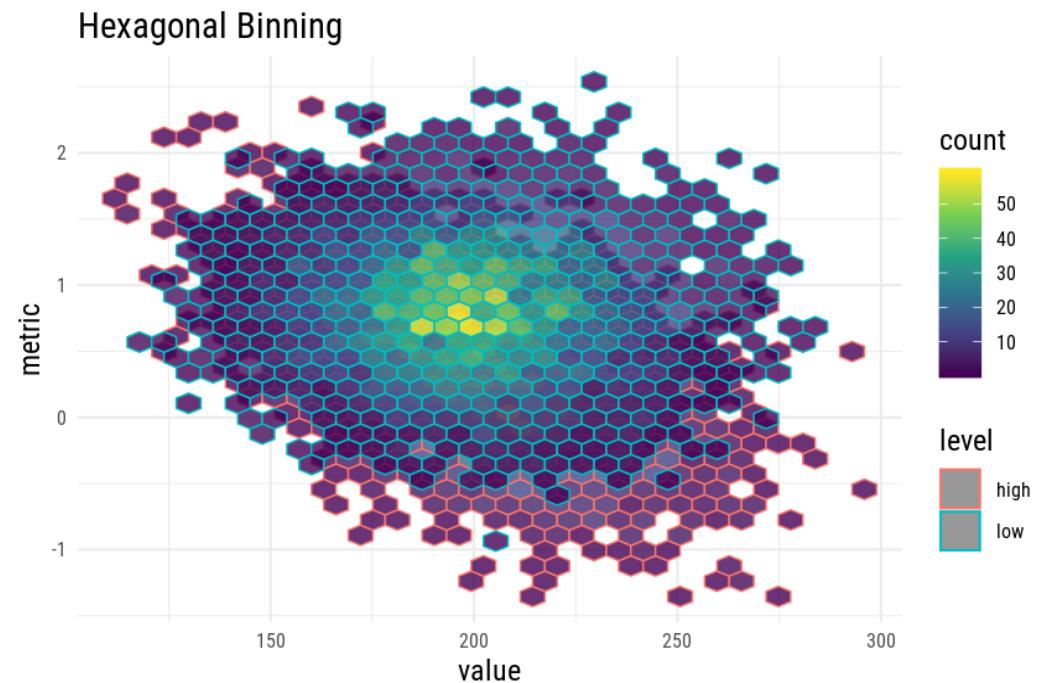
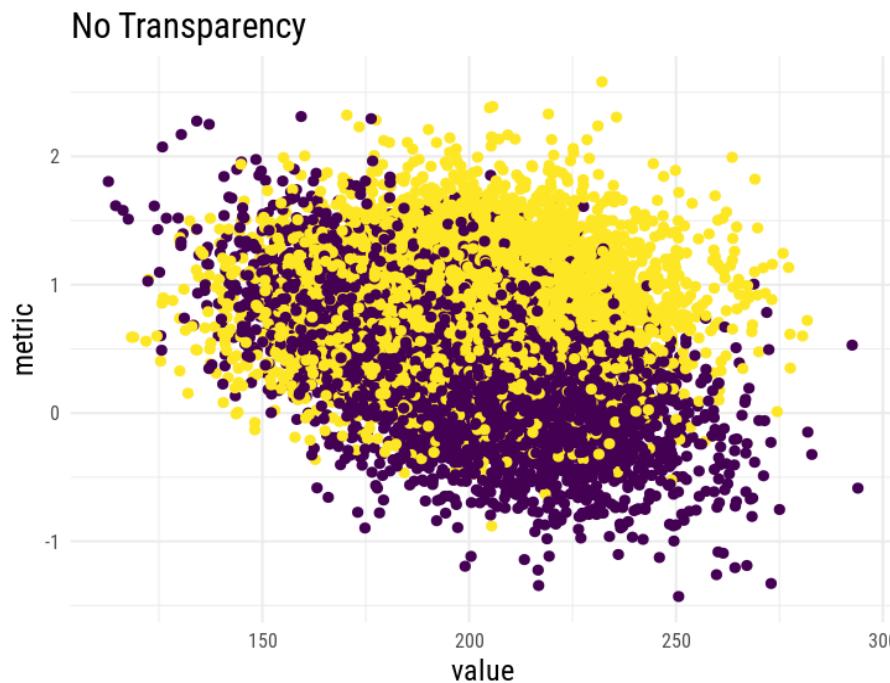
Add **transparency** (alpha). Easy and quick to implement: reveals underlying patterns but hides outliers and sparse data points.



Visualizing High Density Data

Better...

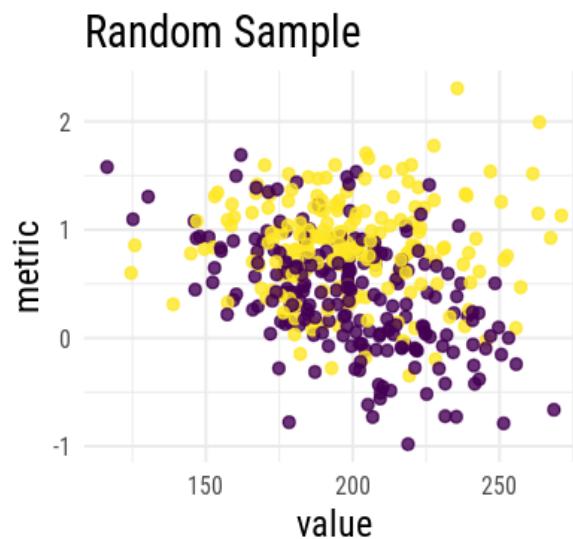
Use **binning techniques** for high density data (they've been around since the mid 80's!). A little harder to read, but won't miss outliers and sparse data.



Visualizing High Density Data

Best...

Critically assess whether you need BIG data in the first place: unless outliers are the main focus of analysis, it's often satisfactory to **sample** the data.



[1] If you are unconvinced, ask yourself: do I really have the whole population despite a huge dataset? Big data provides no protection against bias ↗

Visualizing Changes Over Time

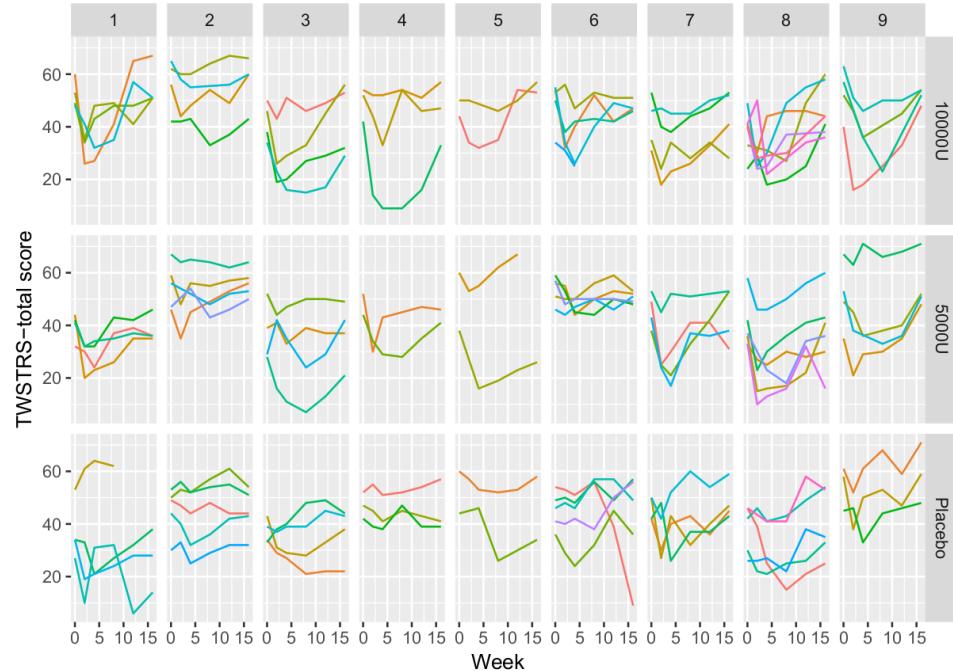
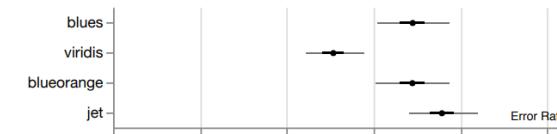
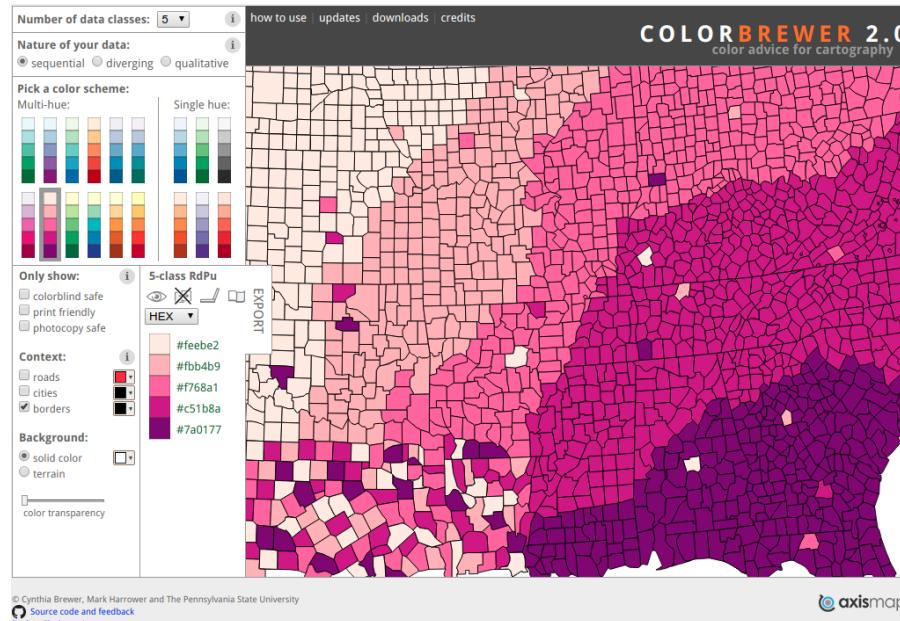


Figure 4.7: Spaghetti plot showing all the raw data on the response variable for each subject, stratified by dose and study site (1–9). Importantly, week 0 (baseline) measurements are included.

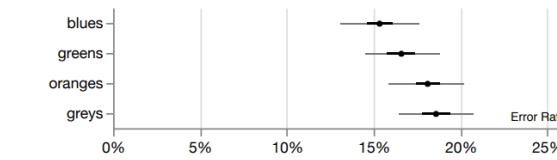
- Works well when the data is not huge.
- Best if stratified.
- x-axis needs not be time, but best if ordinal.

Choosing Appropriate Color

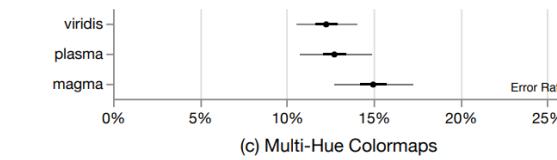
Choice of color is often subject to "*misguided artistry*". There are numerous empirically validated color scales, so the analyst's job is made significantly easier.



(a) Assorted Colormaps



(b) Single-Hue Colormaps



(c) Multi-Hue Colormaps

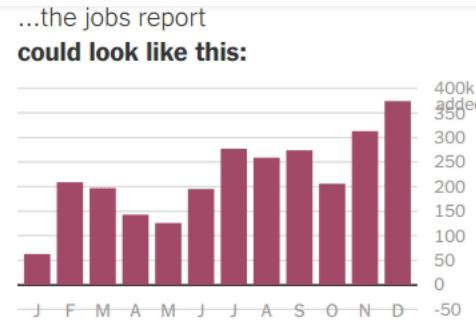
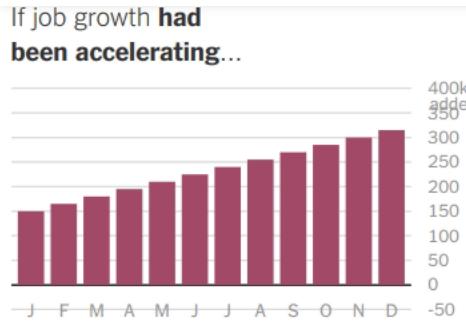
...finally, we confirmed that a rainbow colormap (jet) does indeed perform the worst overall in terms of both time and error, and should be jettisoned.

[1] Brewer. *Colorbrewer 2.0: Color Advice for Cartography* ↗

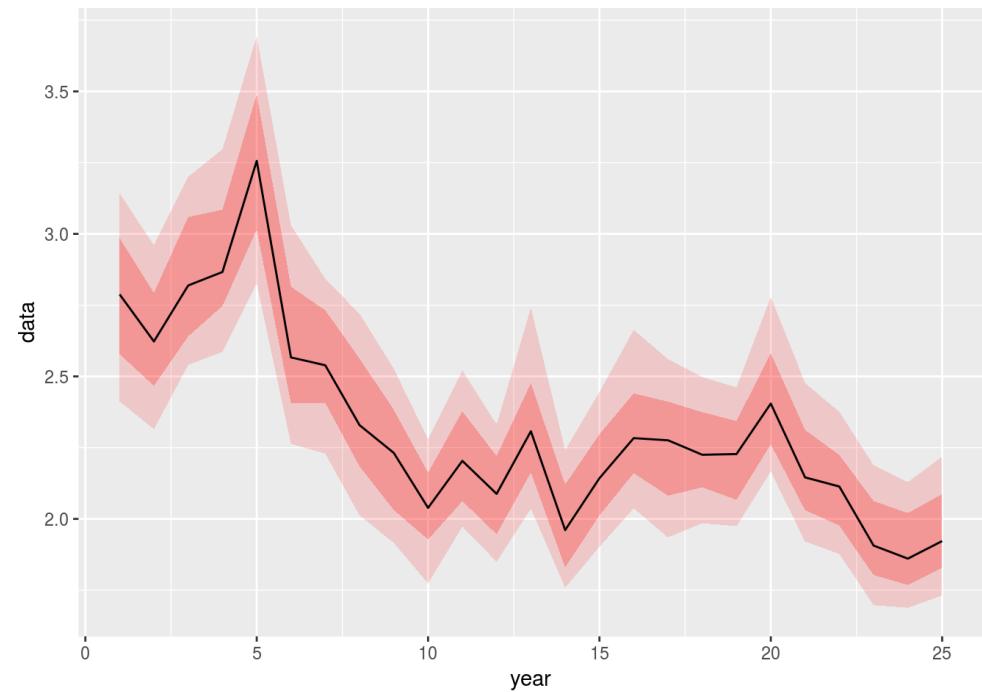
[2] Liu, Heer. *Somewhere Over the Rainbow: An Empirical Assessment of Quantitative Colormaps*. 2018 ↗

Show the Uncertainty

We need to visualize the uncertainty in order to make optimal decisions.

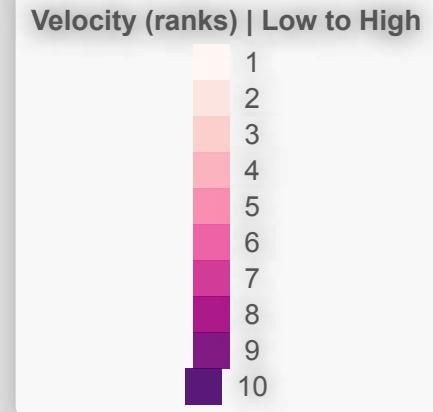
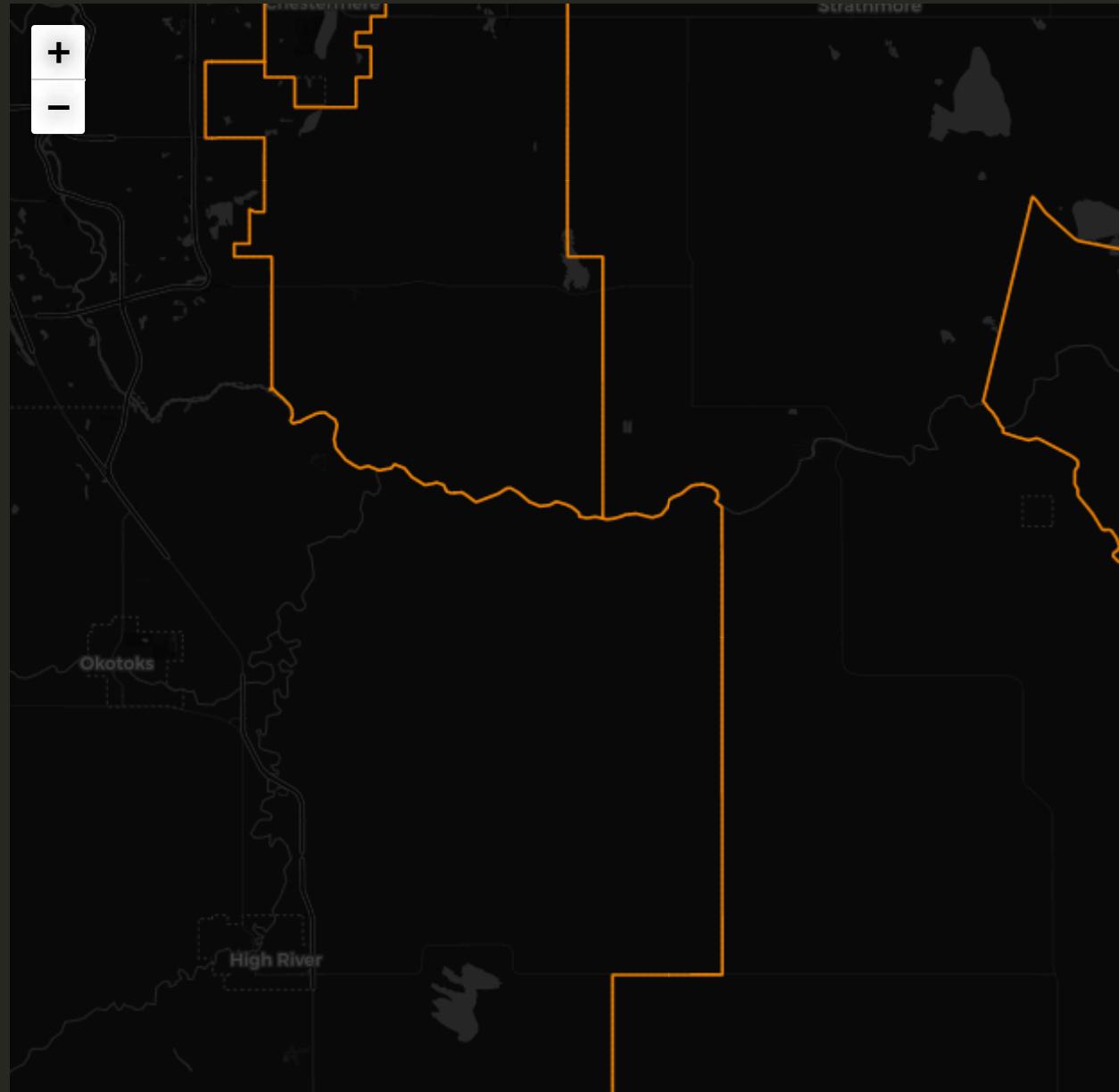


If you squint, you can see evidence of the upward trend, with stronger growth apparent in the second half of the year. But month by month, you wouldn't have any way of knowing if it was a true acceleration, or just a false signal generated by sampling error.



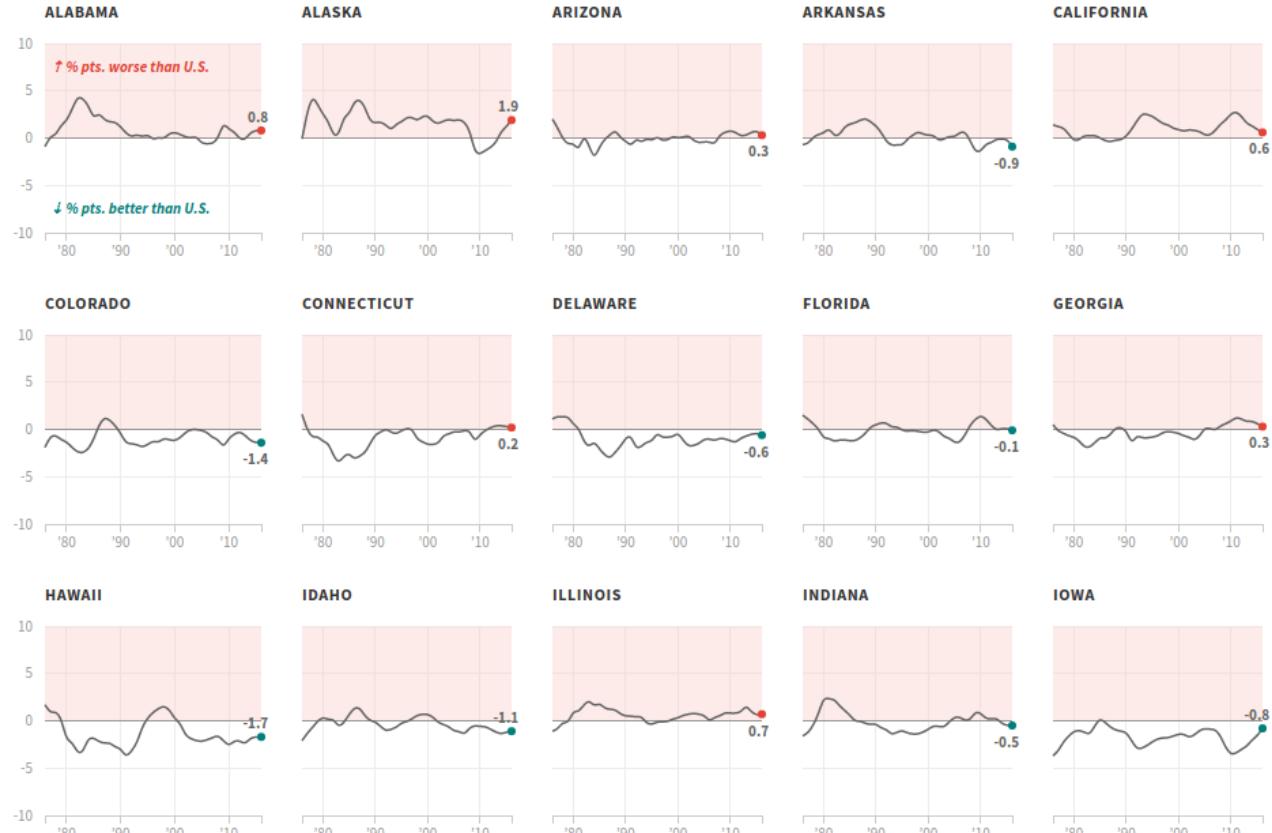
[1] New York Times. How Not to Be Misled by the Jobs Report, 2014 ↗

[2] Chato Osio and Hamon. 2017 ↗



Small Multiples

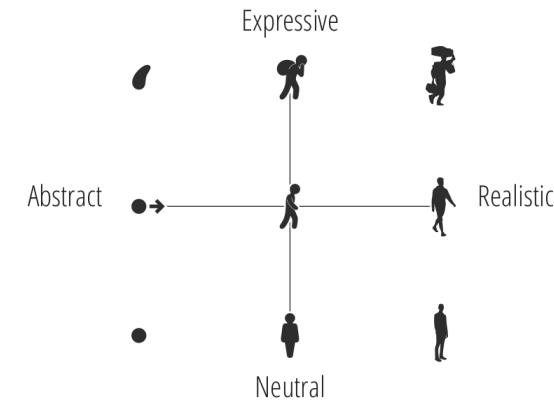
A series of very similar graphs where the eye is drawn on comparing differences among them.



Should We Anthropomorphize Visualizations?

...to attribute human form or personality to things not human

In 2017, Boy et al. (1) conducted visual experiments on 284 participants. They found no evidence that anthropomorphizing visualizations has benefits on empathy.



(1) Boy J. et al. *Showing People Behind Data: Does Anthropomorphizing Visualizations Elicit More Empathy for Human Rights Data?* ↗

7. What to Look for in Visualization Tools? An Opinionated List

What to Look for in Visualization Tools? An Opinionated List

Goal 1: *to identify, highlight features or relationships*

Goal 2: *to summarize, communicate and engage (telling a story)*

Evaluation Item	Notes
Who is it for?	Identify the users and consumers. Likely 2-3 streams.
Goal 1 or goal 2?	Data visualization vs. visual storytelling
No one size fits all	General purpose vs. special purpose
Data preparation approach	Visual composition framework
Data preparation capabilities	Joins, filter, aggregate, sample, rank, reshape, append, grouping, text extraction, logical operators, mathematical expressions
Handling of dates and times	Human readable and adherence to ISO standards.
Time series data preparation	Rolling windows, date padding, durations, rolling statistics, smart aggregations
Breadth of statistics for aggregation	Mean, percentiles, ranks, min/max, counts, totals, conditional statistics, proportions, duration, harmonic mean, geometric mean, sd, IQR...
Missing values	Appropriate handling/warning of missing data
Ability to show uncertainty	Confidence bands, error bars, rangeplots
General purpose smoothers	LOESS, exponential smoothing
Quality of visualizations	Avoids or encourage chartjunk
Breadth of visualizations	Standard graphs and modern graphs (e.g. ridgeline plots, violin plots, circular migration plots, word clouds)

What to Look for in Visualization Tools? An Opinionated List

Evaluation Item	Notes
Learning curve	Effort and time
Community size	Extremely important! Forum, discussion boards, conferences, books, blogs, YouTube, MOOC's
Extensibility	Community contribute added features
Extensibility	Play nice with R, Python?
Connectivity	Beyond common connections. E.g. API's Bloomberg, weather, social media

Don't over-rely on IT advisory and consulting firms.

When evaluating tools, **weighting** and **scoring** tend to be arbitrary and subjective. Scales are also arbitrary with little scientific basis. This usually results in poor choices. Weights are easy to manipulate in order to favour one tool over another. The resulting scoring matrices can be "*worse than useless, leading to worse-than-random decisions*"(1).

Favour a **thorough qualitative evaluation**, ideally from a **pilot** rather than a pseudo-scientific scoring matrix.

[1] Cox LA Jr. What's wrong with risk matrices? Risk Anal. 2008 Apr ↗

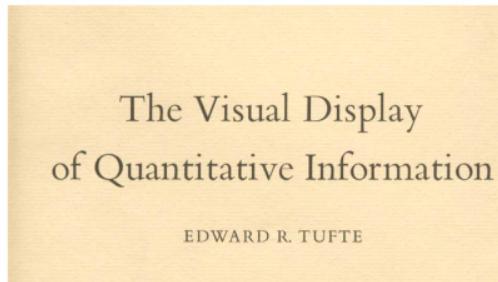
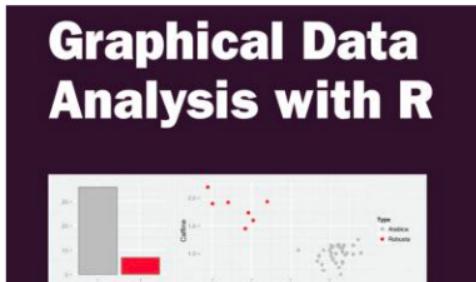
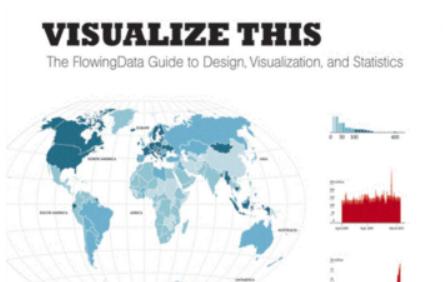
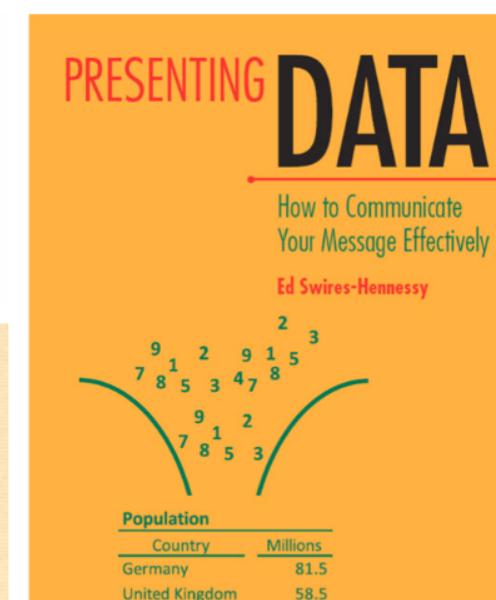
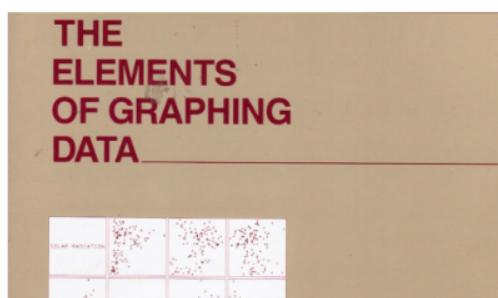
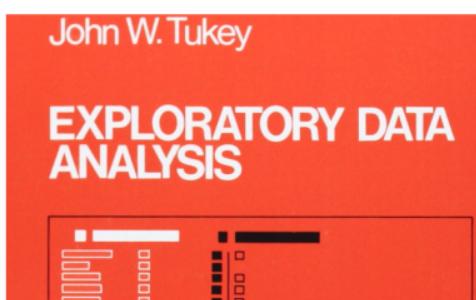
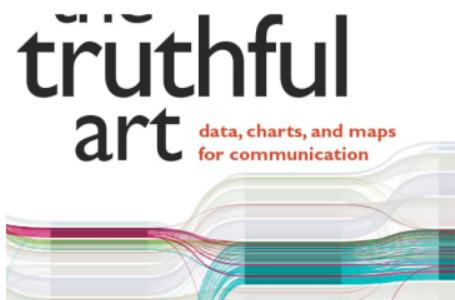
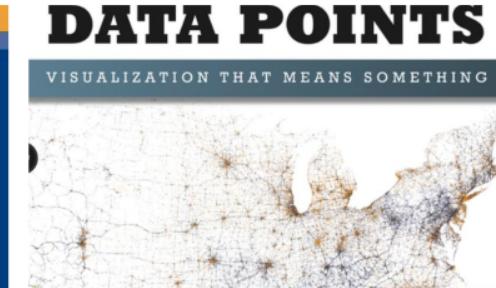
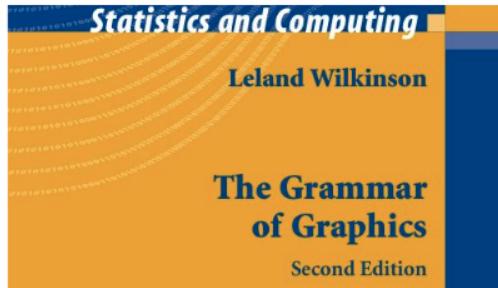
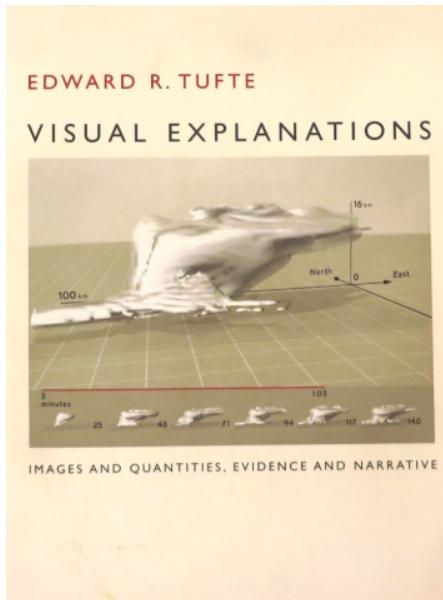
What to Look for in Visualization Tools? An Opinionated List

Remember Tufte: *Numbers become evidence by being in relation to.* Does the tool encourage comparisons? Does it make it easy? A simple test is to plot a stratified box-plot.

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem

-John Tukey

8. References



Author	Title
Tufte, E. R. (1983)	The Visual Display Of Quantitative Information
Tufte, E. R. (1990)	Envisioning Information
Tukey, J. W. (1977)	Exploratory Data Analysis
Few, S. (2012)	Show Me The Numbers. Designing Tables And Graph To Enlighten.
Chiasson T., Gregory D. et al. (2014)	Data + Design
Ehrenberg, A.S.C. (1981)	The Problem Of Numeracy. The American Statistician, Vol. 35, N. 2
Wilkinson, L. (2005)	The Grammar Of Graphics
Simon, H.A. (1996)	The Sciences Of The Artificial.
Cleveland, W.S. (1985)	The Elements Of Graphing Data.
Yau, N. (2011)	Visualize This. The Flowingdata Guide To Design, Visualization And Statistics
Good P.I., Hardin J.W. (2012)	Common Errors In Statistics (And How To Avoid Them)
Gelman A., Unwin A. (2012)	Infovis And Statistical Graphics: Different Goals, Different Looks

Thank you

Thomas Speidel

thomas@speidel.ca ↗

ca.linkedin.com/in/speidel/ ↗

alternative-stats.netlify.com ↗

This presentation and most of the graphs were produced in R), a programming language and software environment for statistical computing and graphics. The Xaringan package was used for the presentation.

