Visualizations in the Analysis Workflow

SPE

Thomas Speidel, P.Stat., Data Scientist

2016/12/12

A copy of this presentation is available on GitHub:
http://

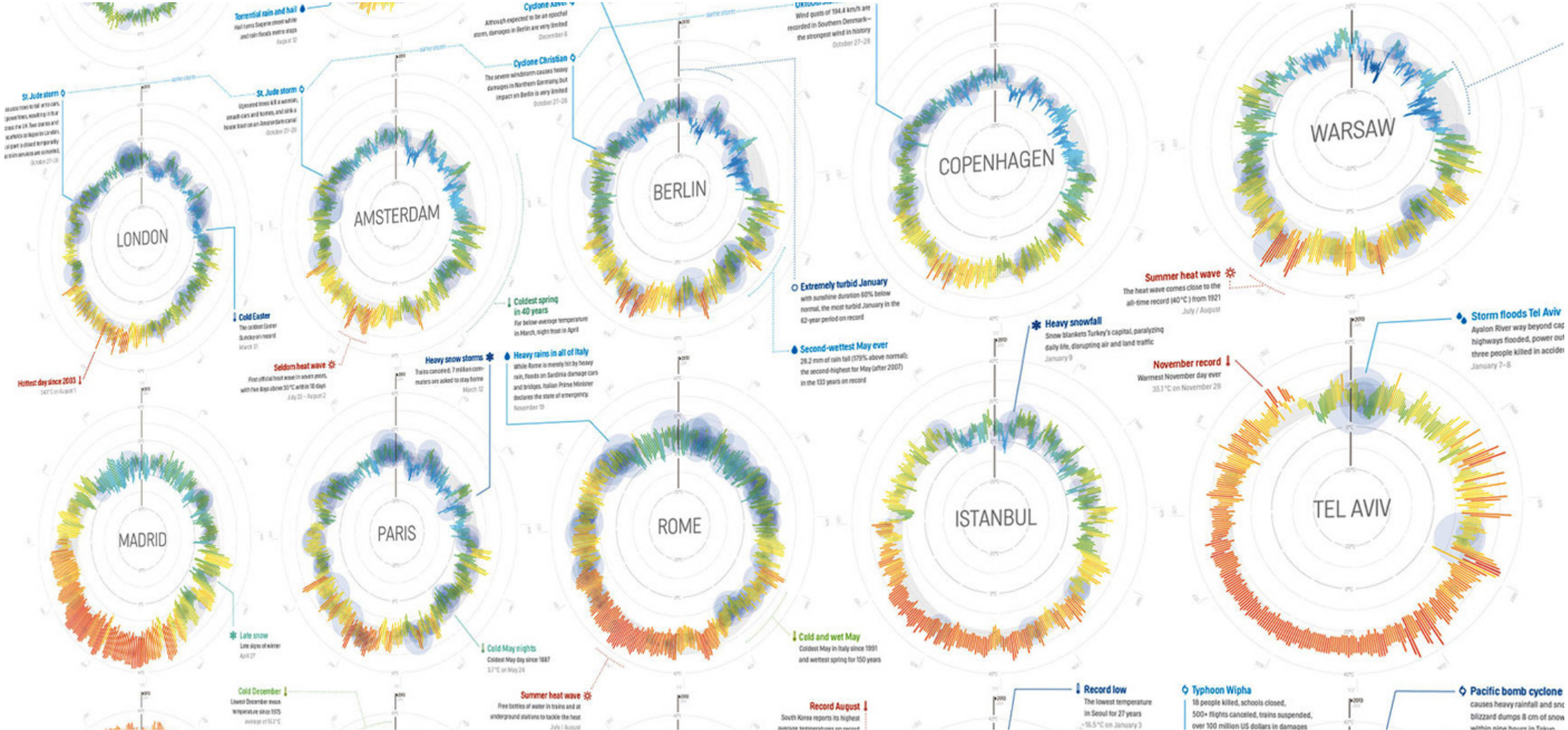
Agenda

1. Visualizations: Objectives and Constraints
2. Univariate Visualizations
3. Bivariate Visualizations
4. Tips & Best Practices
5. References

1. Visualizations: Objectives and Constraints

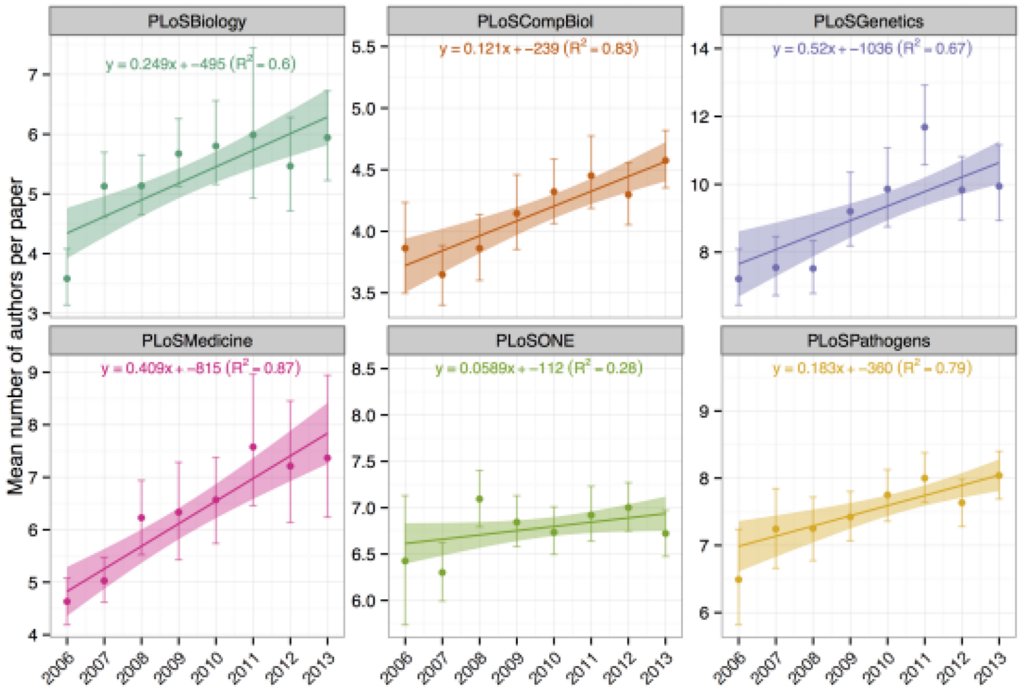
First Goal of Visualizations

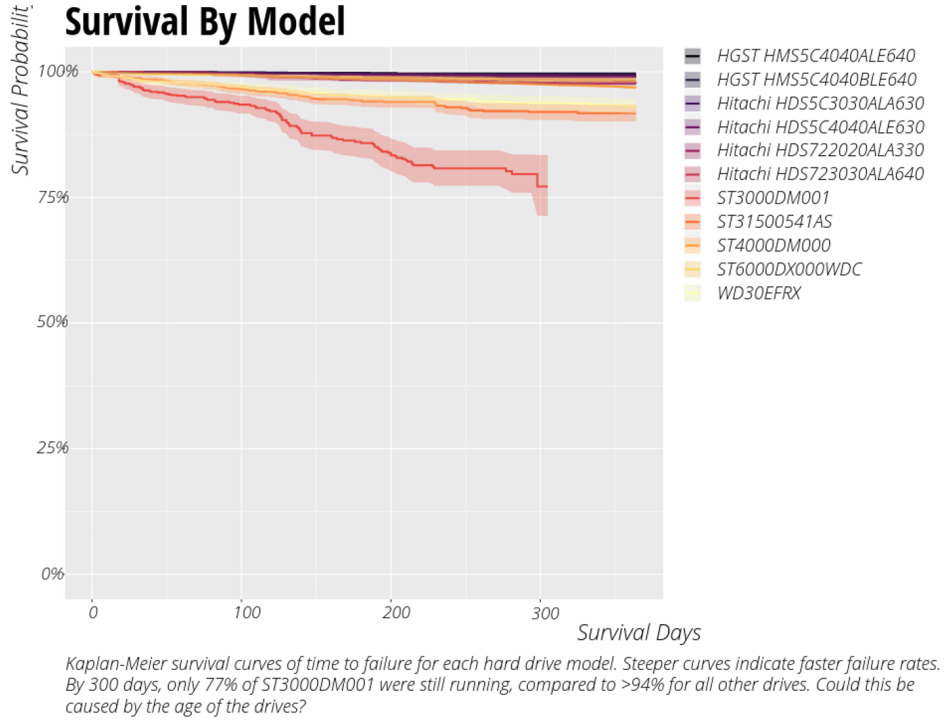
To summarize, communicate and engage (telling a story)



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?

[1] *Gelman, Unwin. 2013*

To Persuade or to Inform?

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

To Persuade

[1] *Spiegelhalter, Pearson, Short. 2011*

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%

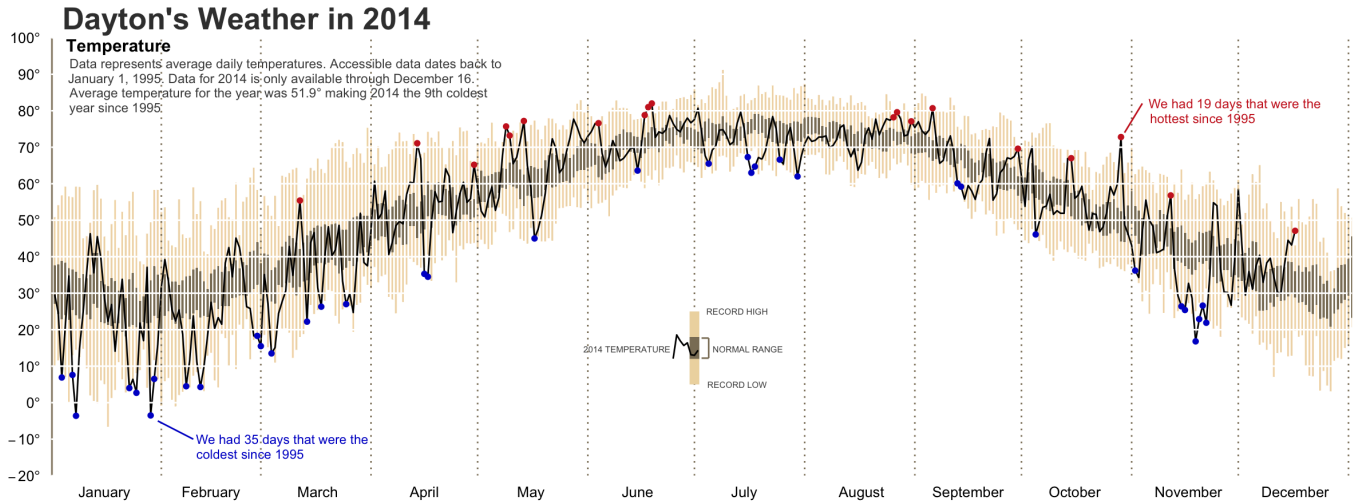
* Percent Daily Values are based on a 2,000 calorie diet. Your Daily Values may be higher or lower depending on your calorie needs.

	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		300g	375g
Dietary Fiber		25g	30g

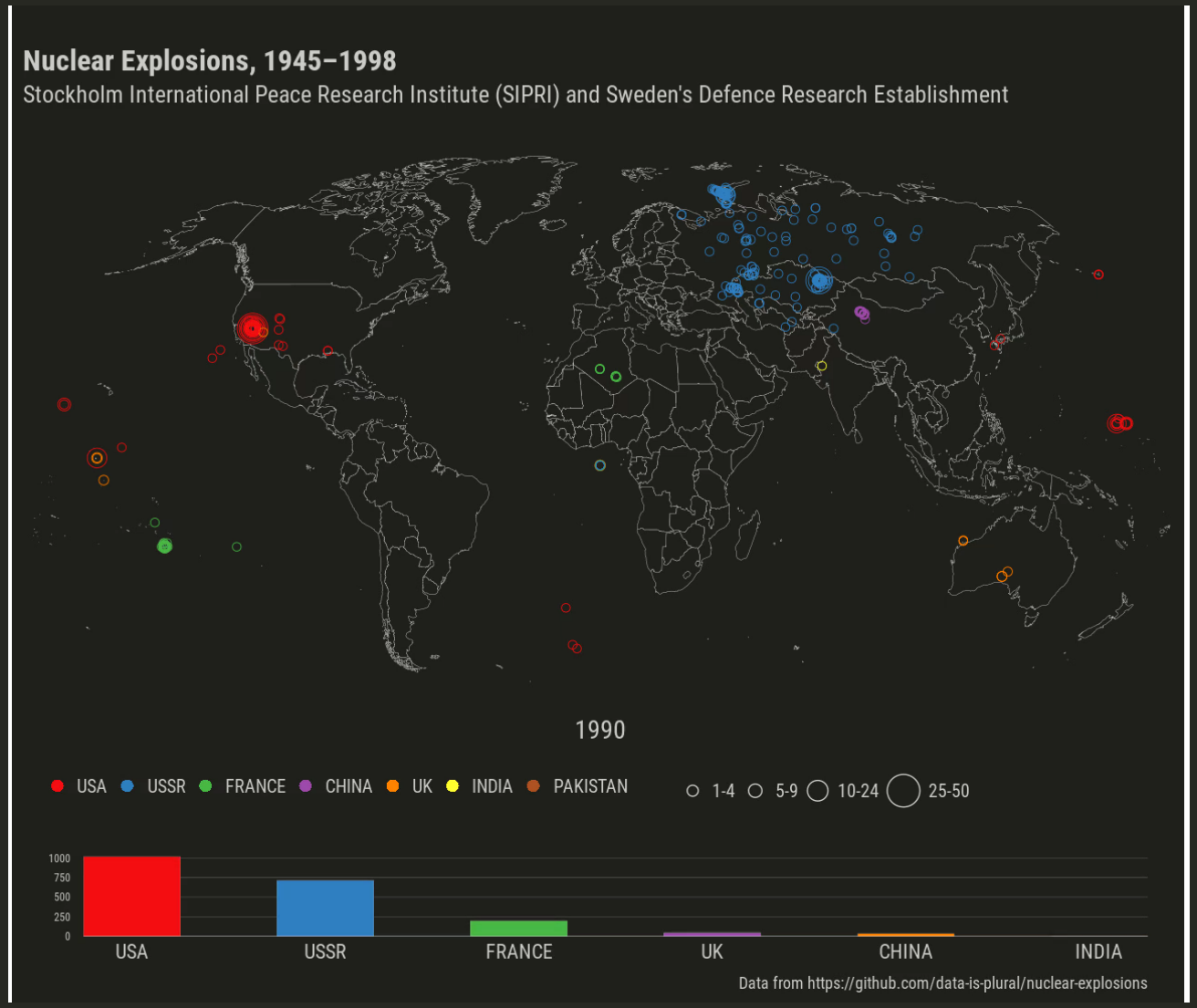
To Remember or to Understand?

It is possible to achieve both **if**: (1):

1. Titles and supporting text should 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*



[1] *Bob Rudis 2016*

2. Univariate Visualizations

Data

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

Exploratory Data Analysis (EDA)

As the analsys 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

[1] *Grolemund, Wickham. R for Data Science, 2017*

Summary Statistics

.

1 Variables **625 Observations**

helium.porosity: Helium Porosity %

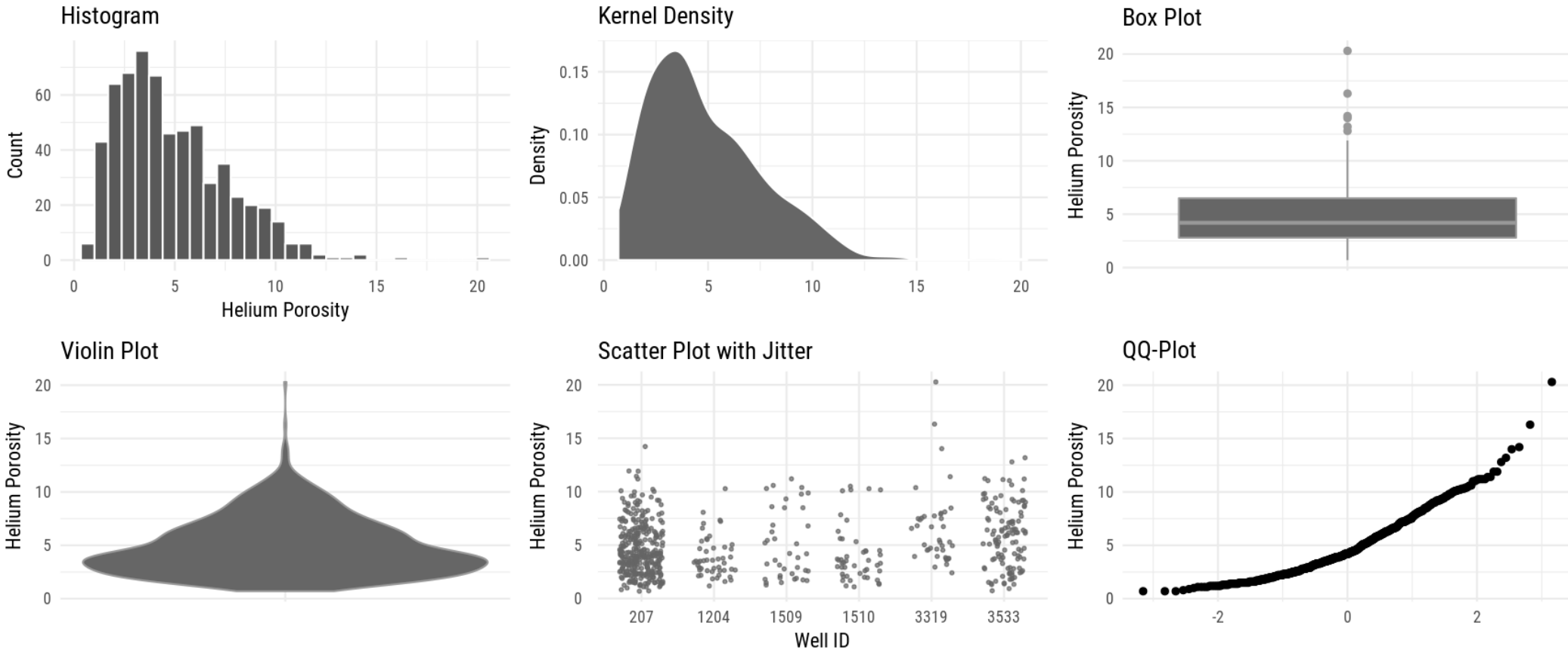
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
625	0	110	1	4.875	2.986	1.5	1.8	2.8	4.2	6.5	8.8	10.0

lowest : 0.7 0.8 0.9 1.0 1.1 , highest: 13.2 14.0 14.2 16.3 20.3

*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



Bivariate

> *Numbers become evidence by being in relation to.*
> -Edward R. Tufte, Visual Explanations: Images & Quantities, Evidence & Narrative (1997)

Tables

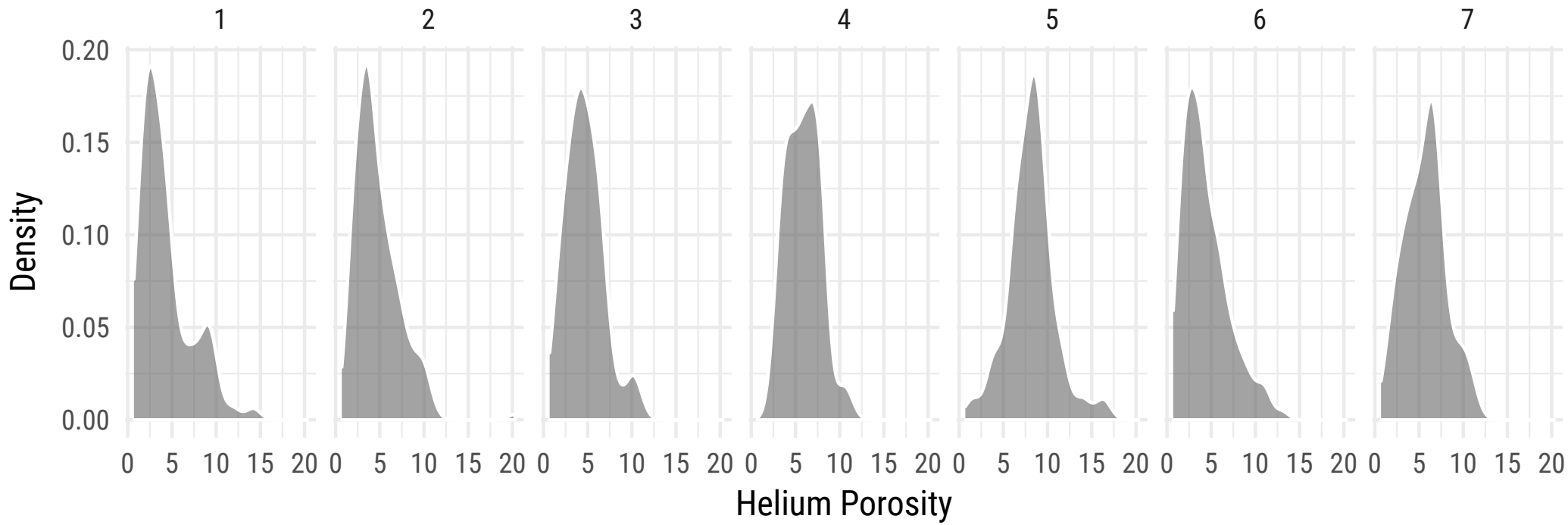
Baseline characteristics by rock type.

		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	ft	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
Density	g/cm3	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 Permeability	md	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 $\overline{X} \pm 1$ SD.
Test used: Kruskal-Wallis test .

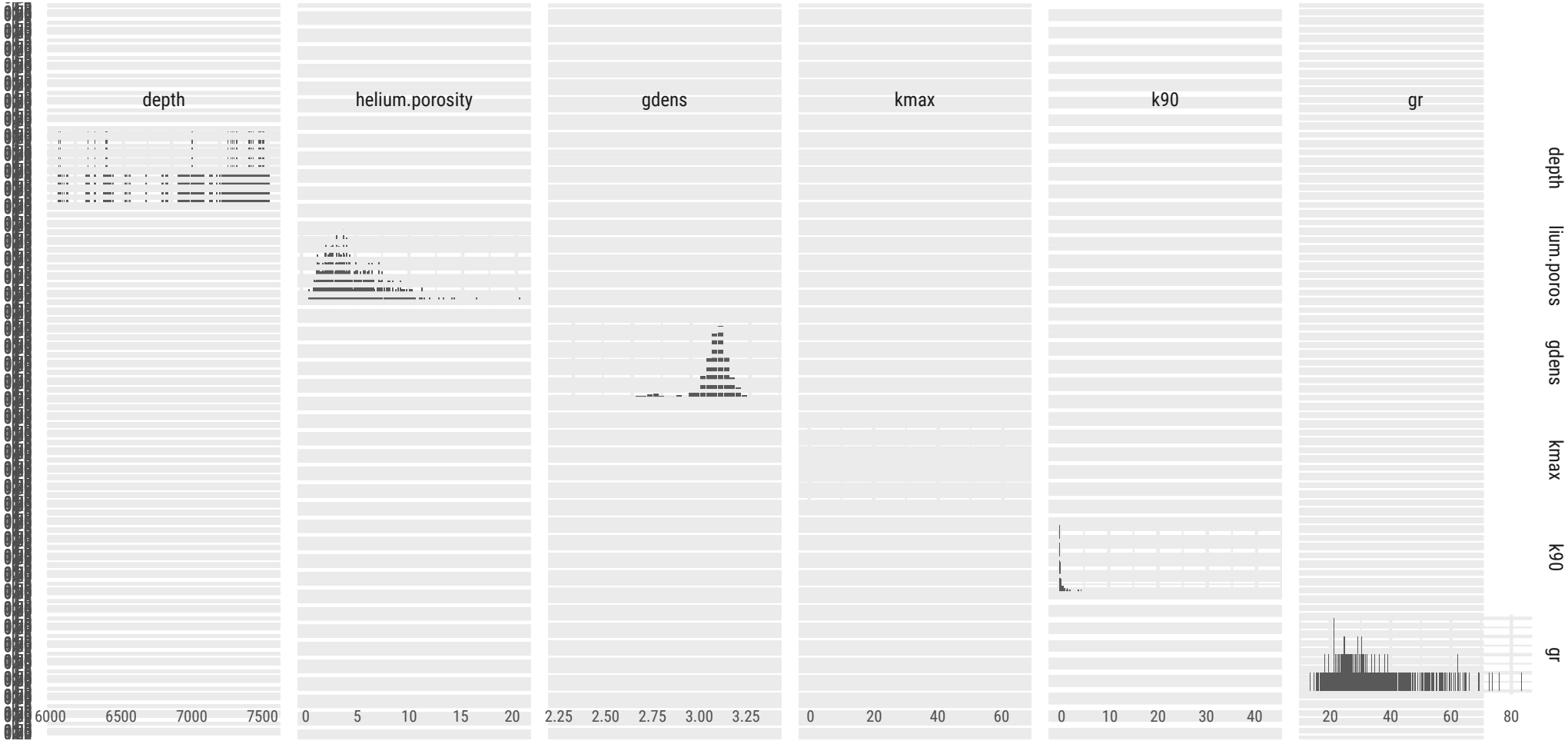
Kernel Density

Kernel density plots have the unique advantage that they are self-standardizing. This make **comparisons** across rock type easier.



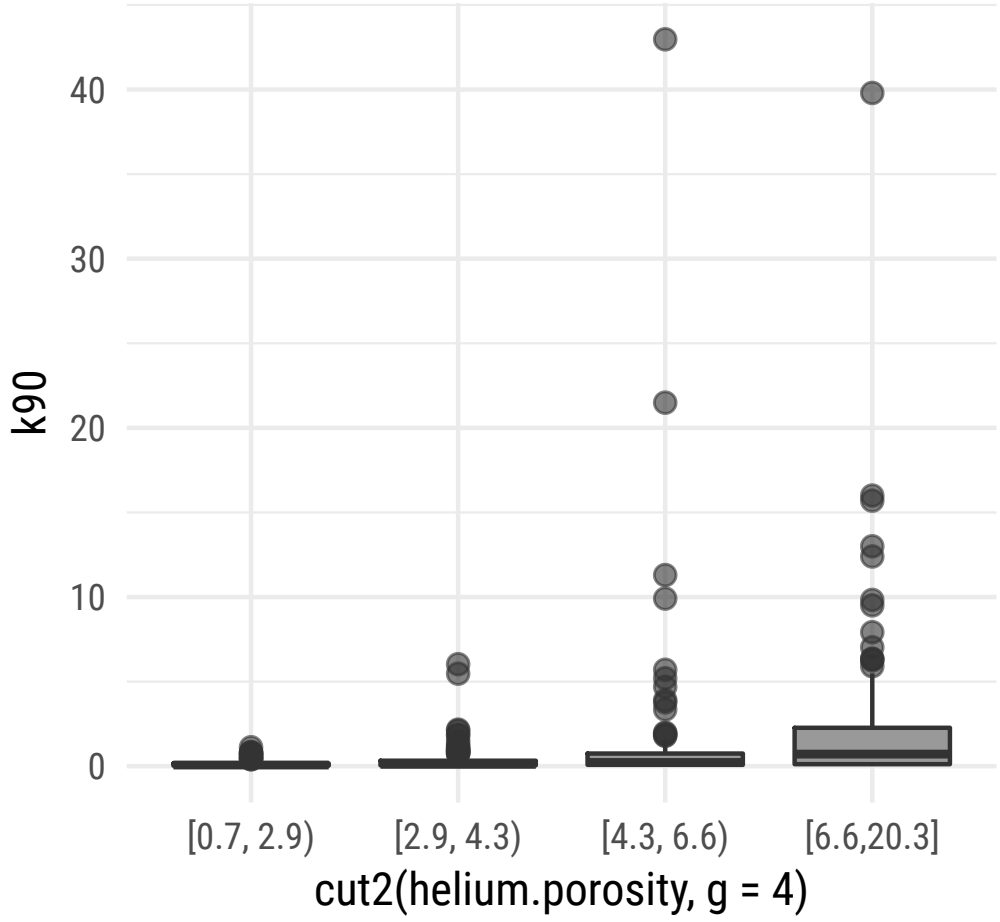
Scatterplot Matrix or Cross Plot

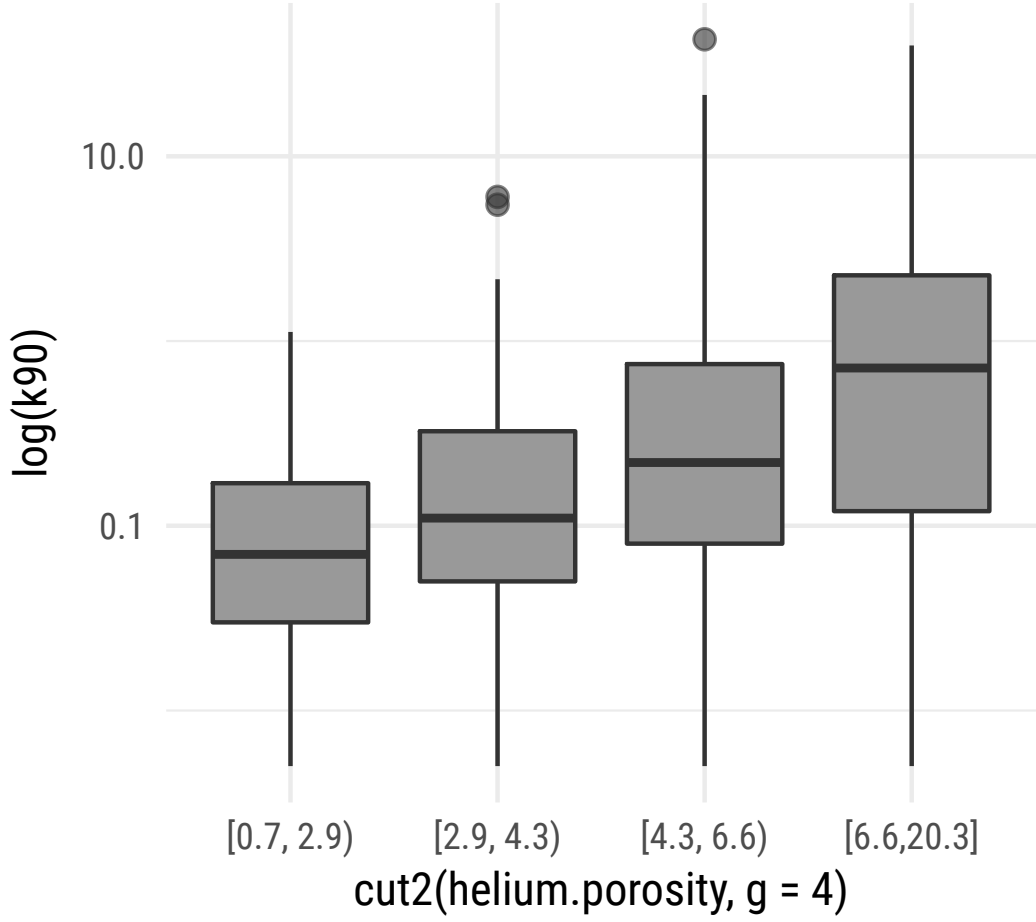
Show all pairwise combinations. Here, I've cheated by removing a few outliers.



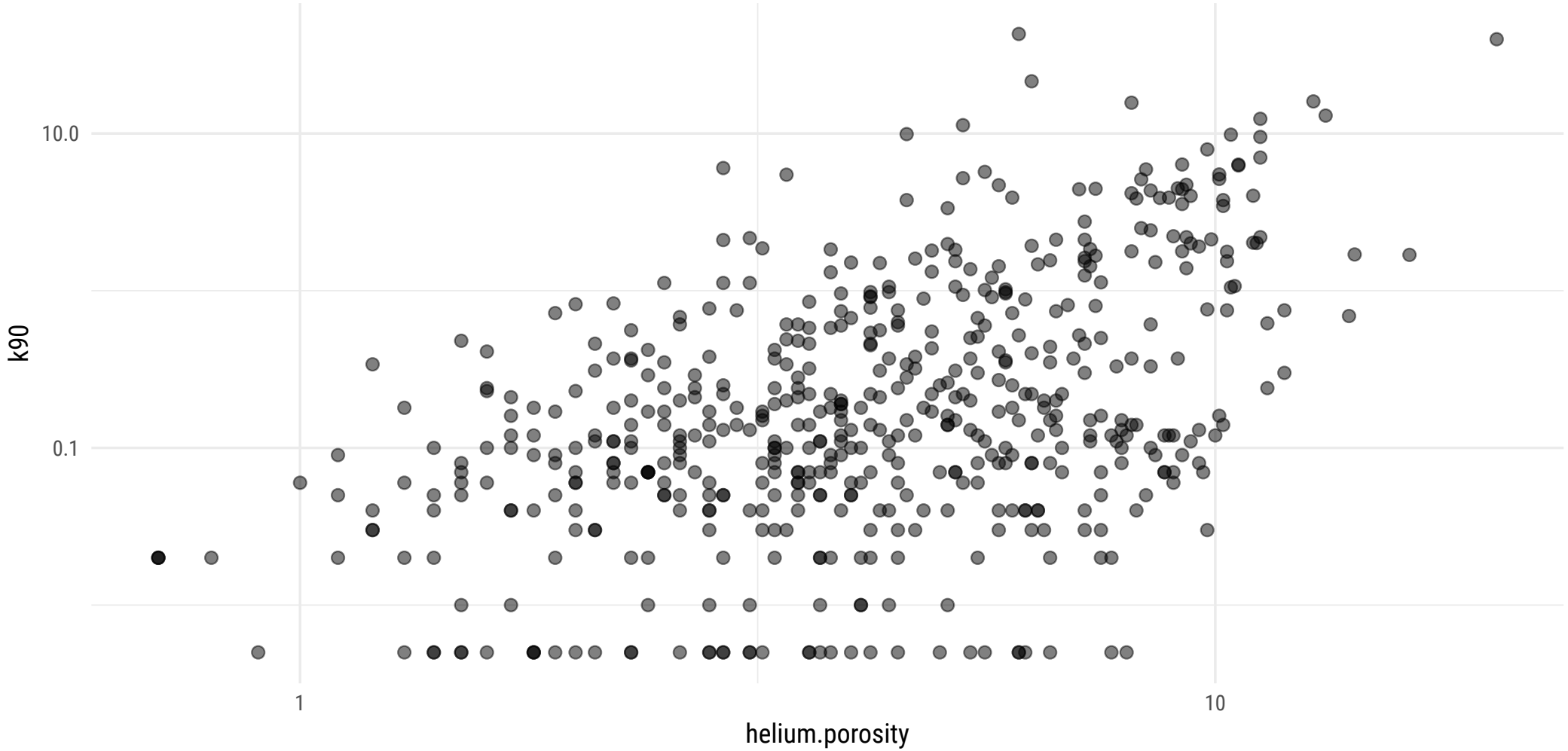
Boxplots

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

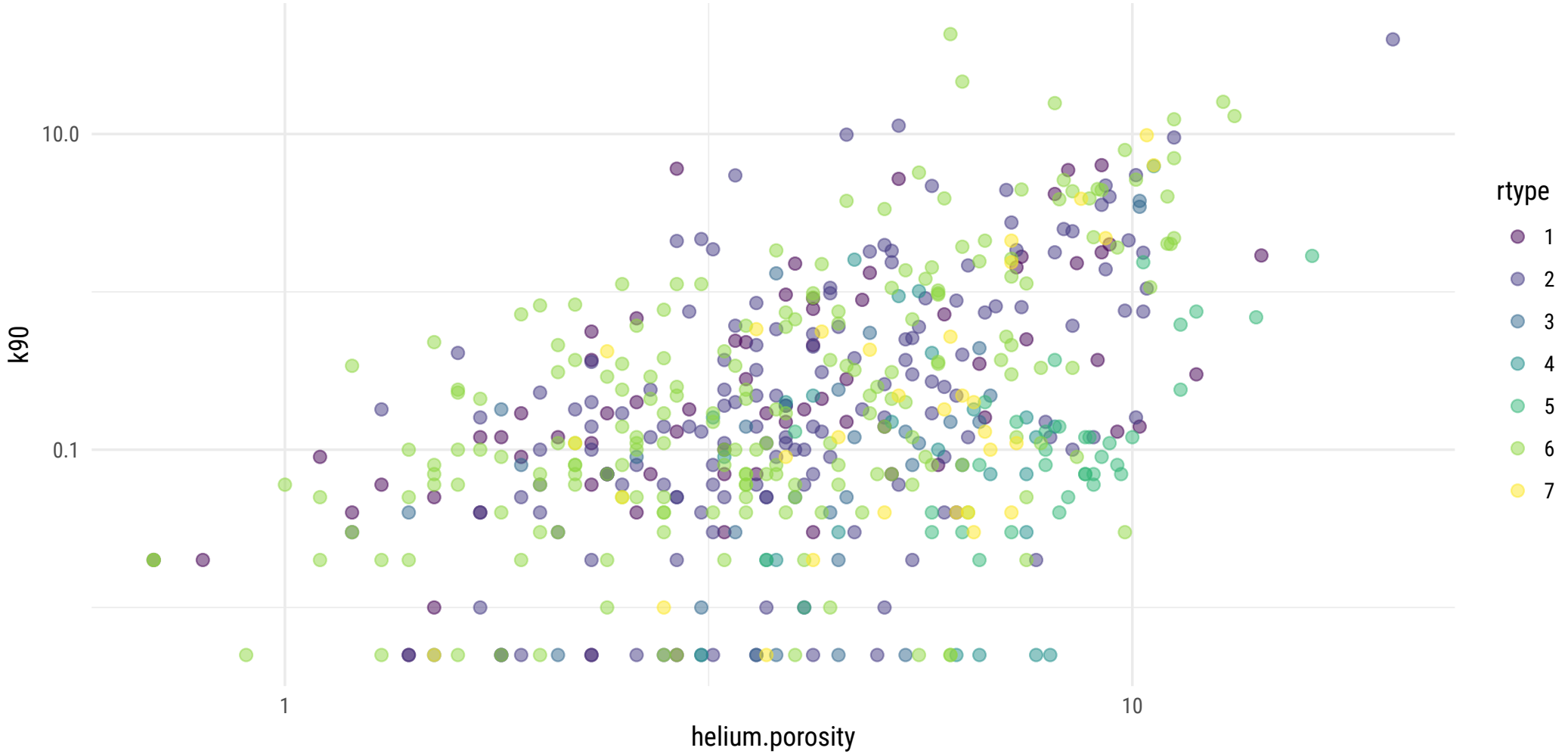




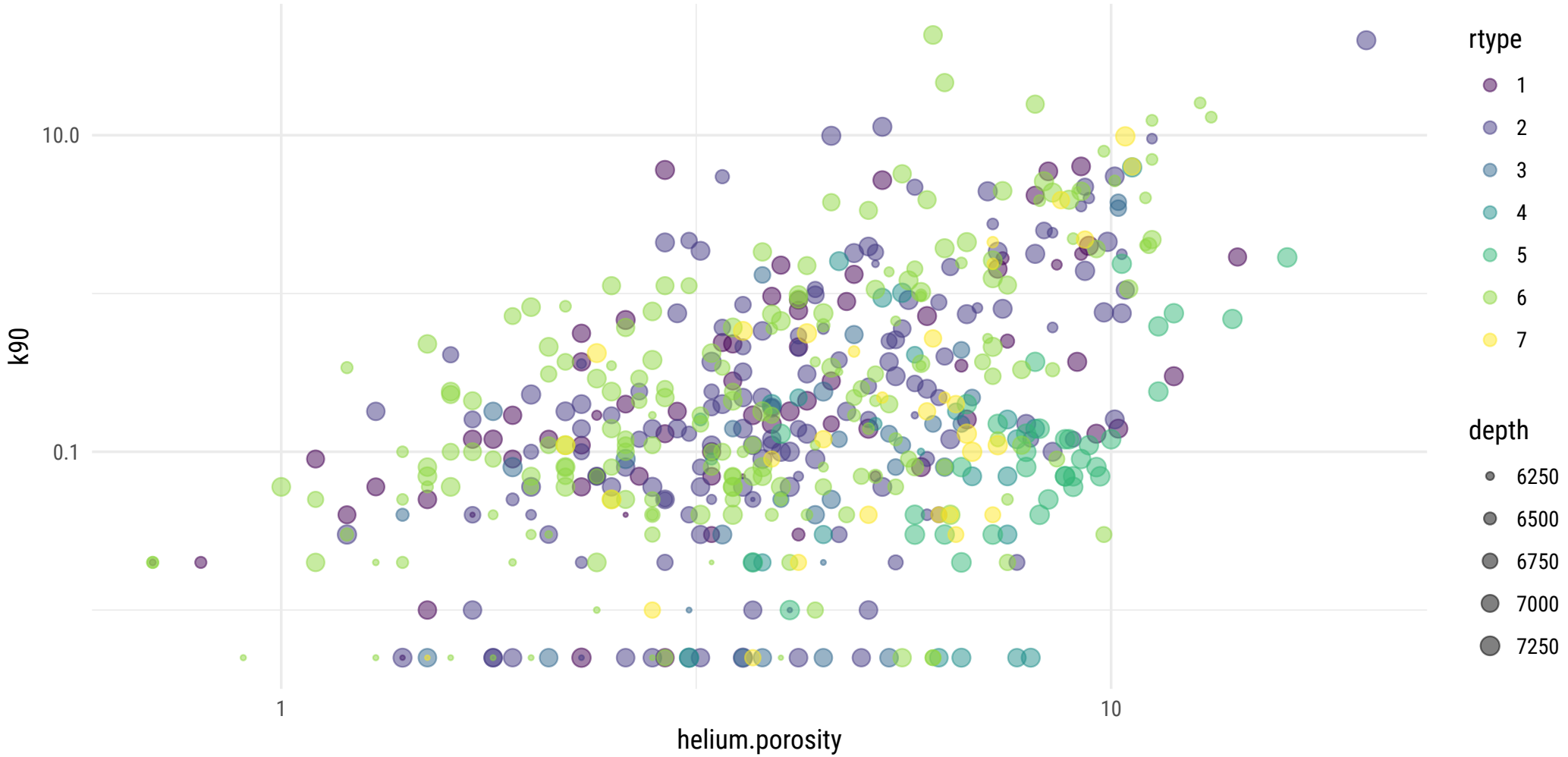
Scatterplots With 2 Variables



Scatterplots With 3 Variables



Scatterplots With 4 Variables



Correlation Matrix

Graphs for Models

It's much easier to interpret a model by graphing it! Compare this:

Logistic (Proportional Odds) Ordinal Regression Model

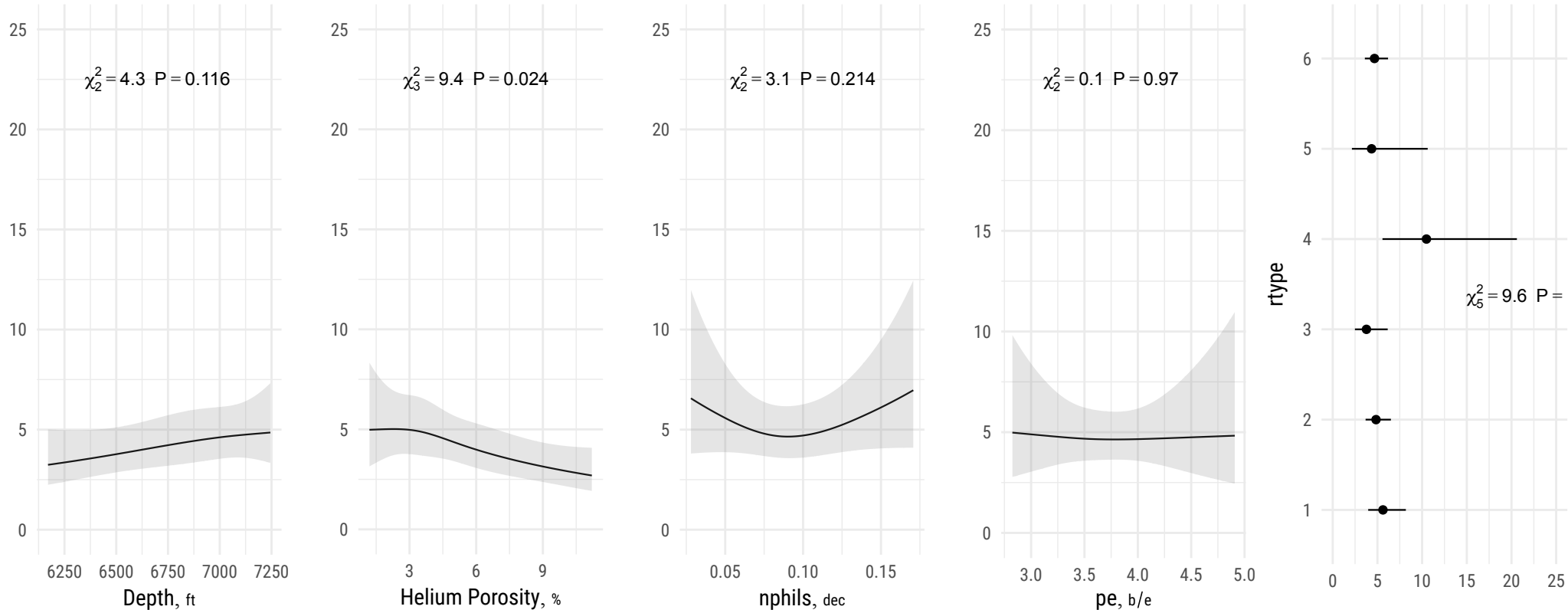
Frequencies of Missing Values Due to Each Variable

```
        k.ratio helium.porosity             pe          nphils
             81               0              0               0
          depth           rtype
              0               0
```

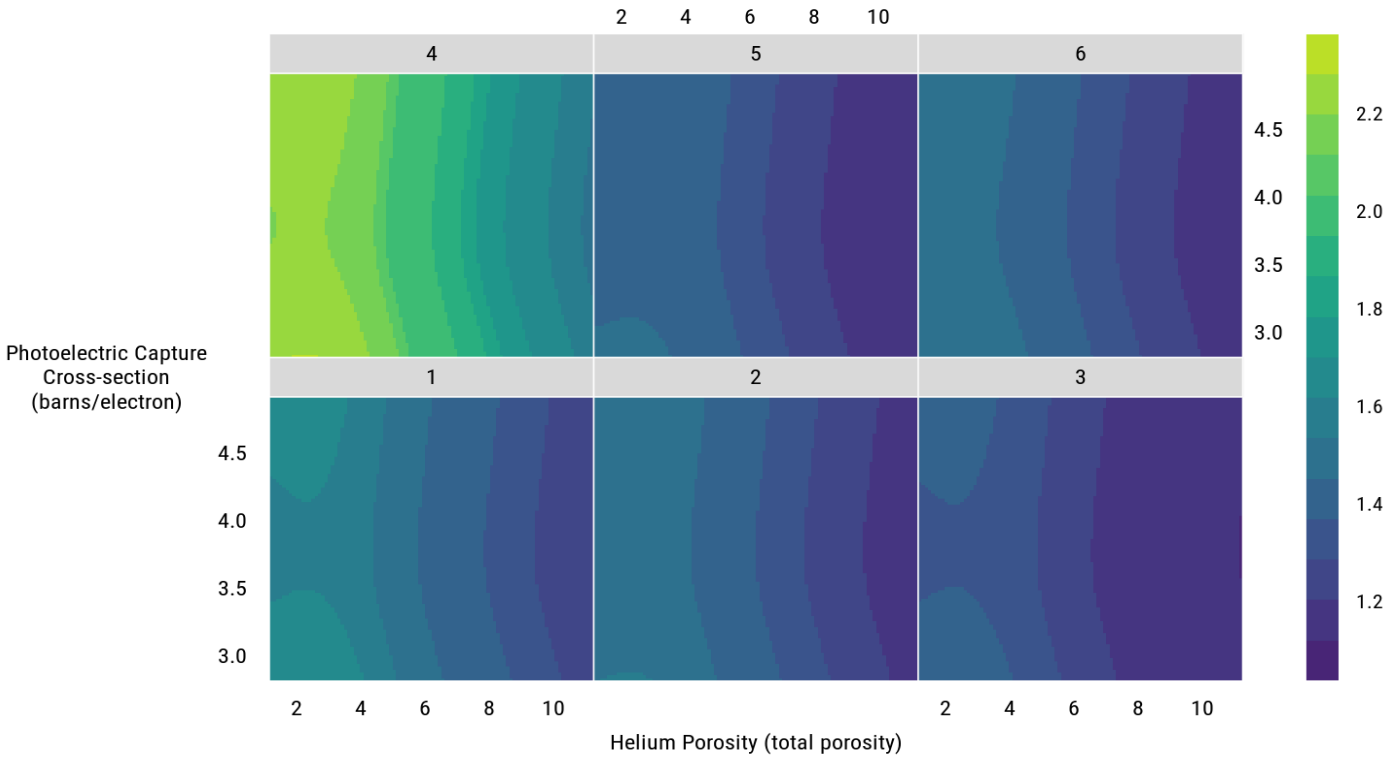
	Model Likelihood Ratio Test	Discrimination Indexes	Rank Discrim. Indexes
Obs 511	LR χ^2 34.84	R^2 0.066	ρ 0.234
Unique Y 318	d.f. 14	g 0.493	
$Y_{0.5}$ 1.433628	Pr(>χ^2) 0.0016	g_r 1.637	
max $\lvert\partial \log L/\partial \beta\rvert$ 1×10^{-6}	Score χ^2 36.21	$\overline{\lvert \Pr(Y \geq \text{median})\text{-}½\rvert}$ 0.083	
	Pr(>χ^2) 0.0010		

	β	S.E.	Wald Z	Pr(>$\lvert Z\rvert$)
helium.porosity	0.012	0.195	0.06	0.9494
helium.porosity'	-0.492	0.965	-0.51	0.6100
helium.porosity''	0.987	2.134	0.46	0.6437
pe	-0.140	0.578	-0.24	0.8086
pe'	0.204	0.906	0.23	0.8214

... with this:


$\chi^2_2 = 4.3$ P = 0.116
Depth, ft
$\chi^2_3 = 9.4$ P = 0.024
Helium Porosity, %
$\chi^2_2 = 3.1$ P = 0.214
nphils, dec
$\chi^2_2 = 0.1$ P = 0.97
pe, b/e
rtype
$\chi^2_5 = 9.6$ P =

... or this:



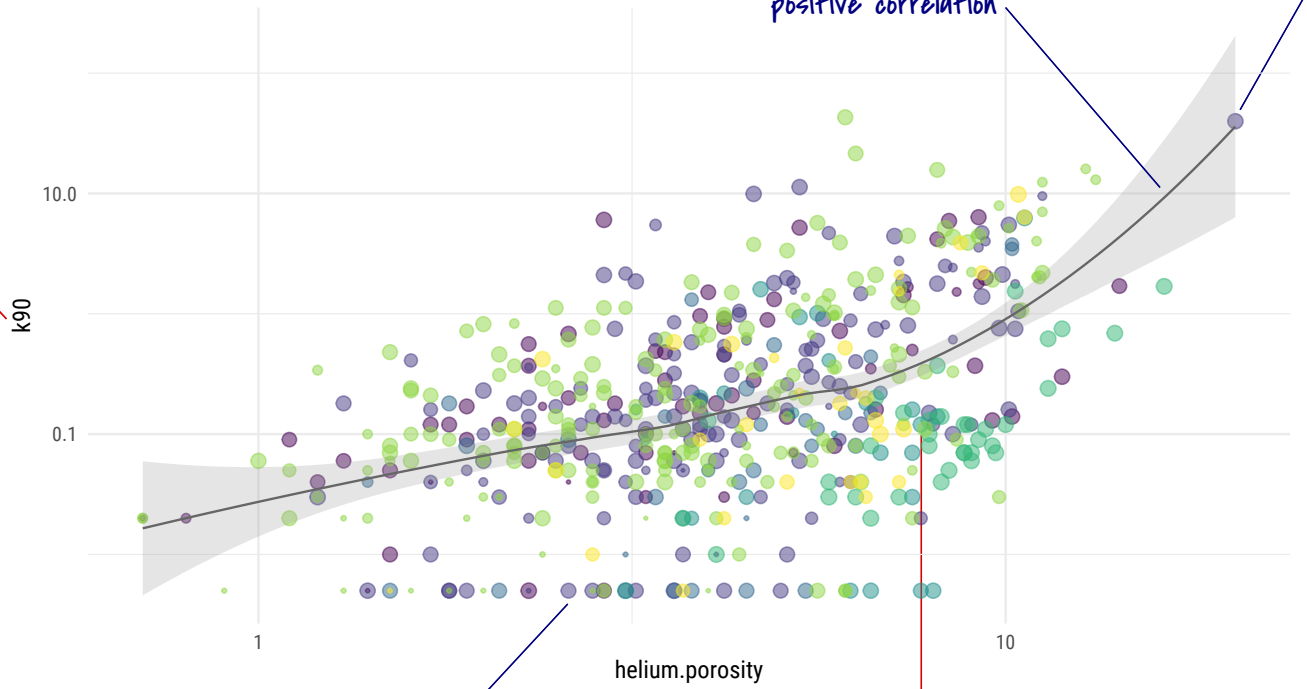
Non-linear regression model showing the simultaneous effect, on the dependent variable permeability ratio (Kmax/k90), of changing two continuous variables (helium porosity and photoelectric capture cross-section log) while holding (marginalizing) all other variables fixed at their median.

4. 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

Overall trend line suggests positive correlation

Outlier?

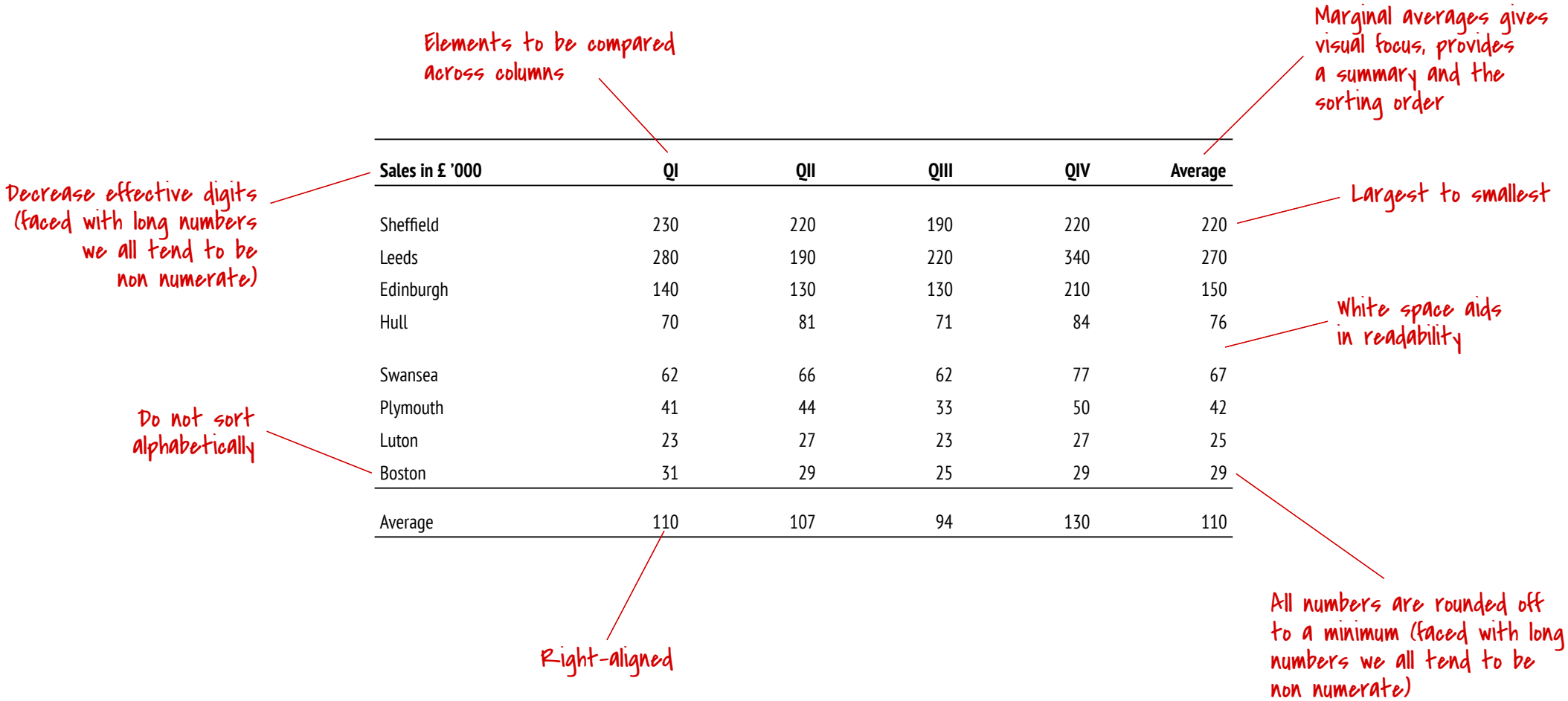
Use of colors:
1. Wide gamut
2. Perceptually uniform
3. Colorblind friendly
(e.g. colorbrewer, viridis)

Flooring effect?
Lower detectable limit?

Use transparency for high density (alpha)

Also called bubble charts.
Use cautiously. Some viz tools get this wrong! Scaling should be $r = (z / pi)^{0.5}$ where z=scaling variable

Table Tips



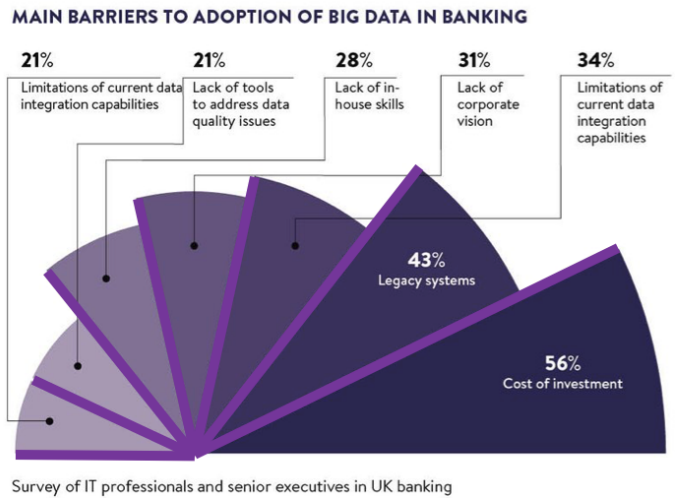
Sales in £ '000	QI	QII	QIII	QIV	Average
Sheffield	230	220	190	220	220
Leeds	280	190	220	340	270
Edinburgh	140	130	130	210	150
Hull	70	81	71	84	76
Swansea	62	66	62	77	67
Plymouth	41	44	33	50	42
Luton	23	27	23	27	25
Boston	31	29	25	29	29
Average	110	107	94	130	110

[1] *Recreated from Ehrenberg, 1981*

Data-to-Ink Ratio

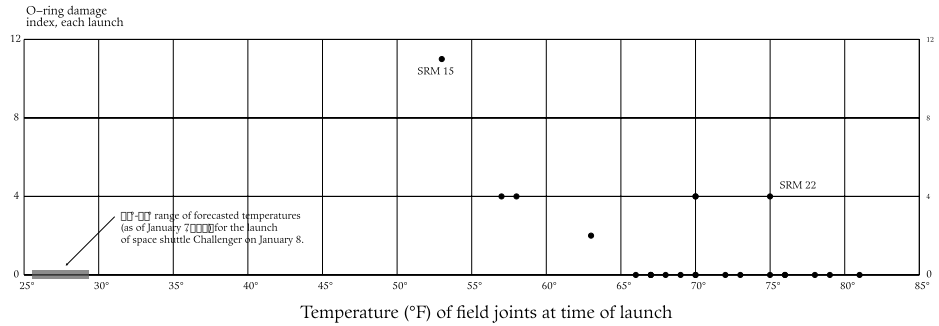
Chartjunk

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

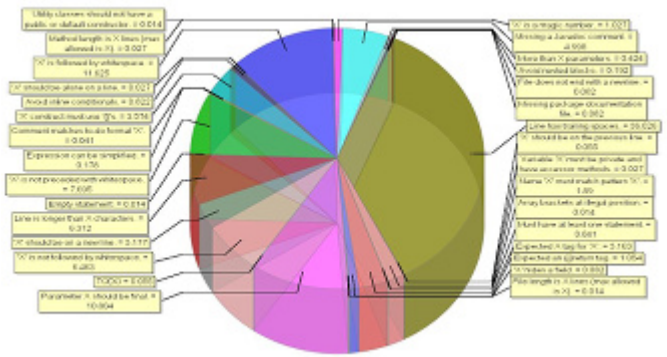
Of all the ink used in a graph, the amount of ink that **transports information**. We want this to be high. (Tufte, 1983, 2006)



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 ½ or ¼ 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

(1)

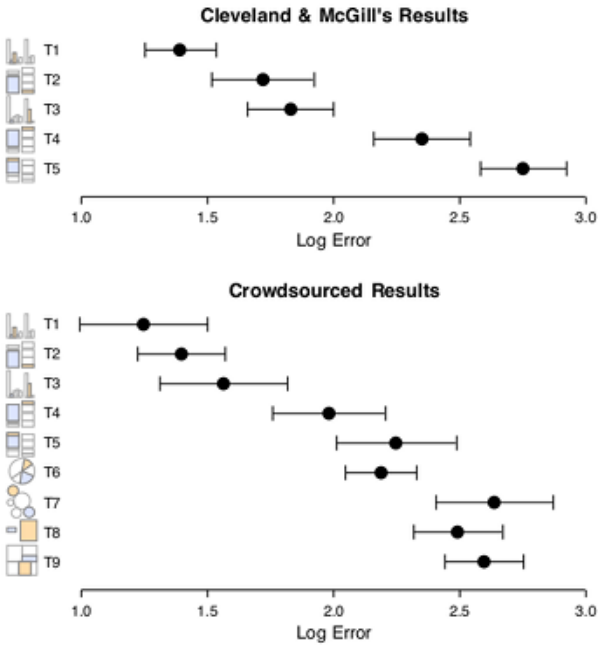


Figure 4: Proportional judgment results (Exp. 1A & B). Top: Cleveland & McGill's [7] lab study. Bottom: MTurk studies. Error bars indicate 95% confidence intervals.

Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods

WILLIAM S. CLEVELAND and ROBERT McGILL*

The subject of graphical methods for data analysis and for data presentation needs a scientific foundation. In this article we take a few steps in the direction of establishing such a foundation. Our approach is based on *graphical perception*—the visual decoding of information encoded on graphs—and it includes both theory and experimentation to test the theory. The theory deals with a small but important piece of the whole process of graphical perception. The first part is an identification of a set of *elementary perceptual tasks* that are carried out when people extract quantitative information from graphs. The second part is an ordering of the tasks on the basis of how accurately people perform them. Elements of the theory are tested by experimentation in which subjects record their judgments of the quantitative information on graphs. The experiments validate these elements but also suggest that the set of elementary tasks should be expanded. The theory provides a guideline for graph construction: Graphs should employ elementary tasks as high in the ordering as possible. This principle is applied to a variety of graphs, including bar charts, divided bar charts, pie charts, and statistical maps with shading. The conclusion is that radical surgery on these popular graphs is needed, and as replacements we offer alternative graphical forms—*dot charts*, *dot charts with grouping*, and *framed-rectangle charts*.

KEY WORDS: Computer graphics; Psychophysics.

1. INTRODUCTION

Nearly 200 years ago William Playfair (1786) began the serious use of graphs for looking at data. More than 50 years ago a battle raged on the pages of the *Journal of the American Statistical Association* about the relative merits of bar charts and pie charts (Eells 1926; Croxton 1927; Croxton and Stryker 1927; von Huhn 1927). Today graphs are a vital part of statistical data analysis and a vital part of communication in science and technology, business, education, and the mass media.

Still, graph design for data analysis and presentation is largely unscientific. This is why Cox (1978) argued, "There is a major need for a theory of graphical methods" (p. 5), and why Kruskal (1975) stated "in choosing, constructing, and comparing graphical methods we have little to go on but intuition, rule of thumb, and a kind of master-to-apprentice passing along of information. . . . there is neither theory nor systematic body of experiment as a guide" (p. 28–29).

There is, of course, much good common sense about how to make a graph. There are many treatises on graph construction (e.g., Schmid and Schmid 1979), bad practice has been uncovered (e.g., Tufte 1983), graphic designers certainly have shown us how to make a graph appealing to the eye (e.g., Marcus et al. 1980), statisticians have thought intensely about graphical methods for data analysis (e.g., Tukey 1977; Chambers et al. 1983), and cartographers have devoted great energy to the construction of statistical maps (Bertin 1973; Robinson, Sale, and Morrison 1978). The ANSI manual on time series charts (American National Standards Institute 1979) provides guidelines for making graphs, but the manual admits, "This standard . . . sets forth the best current usage, and offers standards 'by general agreement' rather than 'by scientific test'" (p. iii).

In this article we approach the science of graphs through human graphical perception. Our approach includes both theory and experimentation to test it.

The first part of the theory is a list of elementary perceptual tasks that people perform in extracting quantitative information from graphs. In the second part we hypothesize an ordering of the elementary tasks based on how accurately people perform them. We do not argue that this accuracy of quantitative extraction is the only aspect of a graph for which one might want to develop a theory, but it is an important one.

The theory is testable; we use it to predict the relative performance of competing graphs, and then we run experiments to check the actual performance. The experiments are of two types: In one, once the graphs are drawn, the evidence appears so strong that it is taken prima facie to have established the case. When a strong effect is perceived by the authors' eyes and brains, it is likely that it will appear to most other people as well. In

* William S. Cleveland and Robert McGill are statisticians at AT&T Bell Laboratories, Murray Hill, NJ 07974. The authors are indebted to John Chambers, Ram Gnanadesikan, David Krantz, William Kruskal, Colin Mallows, Frederick Mosteller, Henry Pollak, Paul Tukey, and the JASA reviewers for important comments on an earlier version of this article.

© Journal of the American Statistical Association
September 1984, Volume 79, Number 387
Applications Section

Cleveland and McGill, 1984 and Heer and Bostok, 2010*

Avoid 3D

Why

Removes clarity, adds obfuscation, inhibits comprehension, does not help retain information, third dimension is usually inexistent (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.

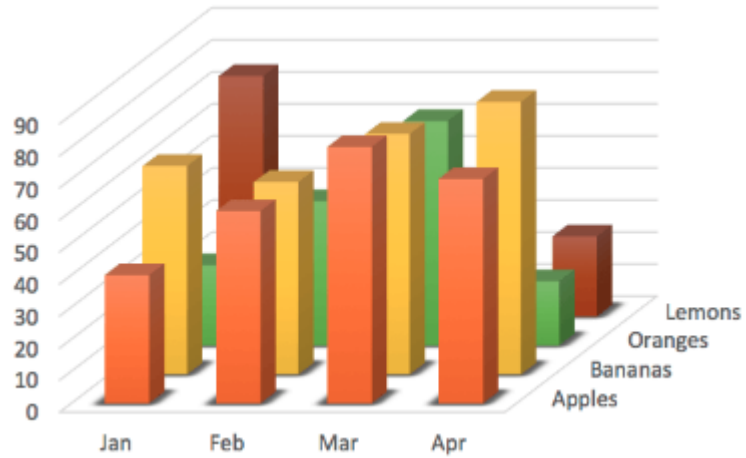
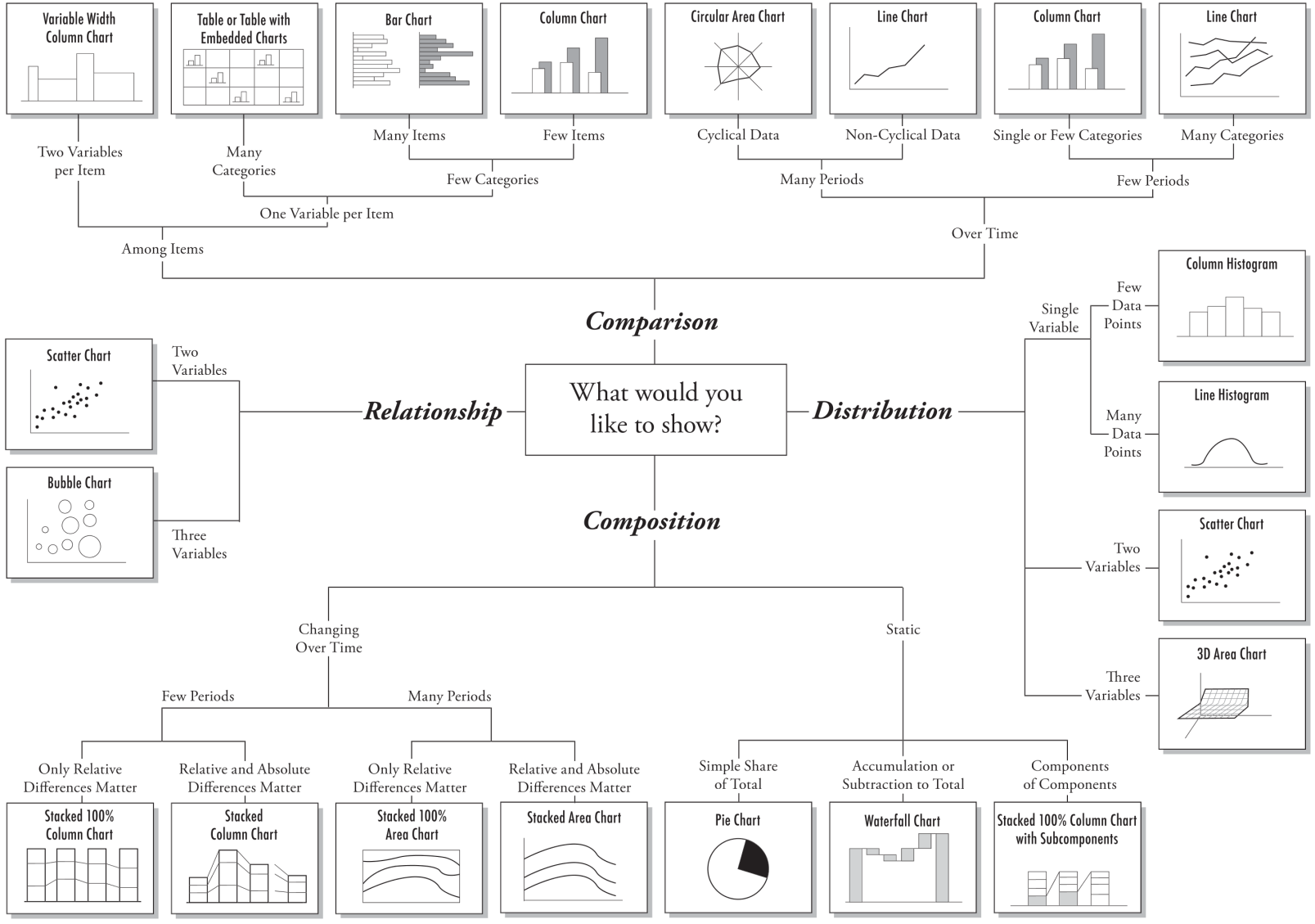


Chart Suggestions—A Thought-Starter



> *An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem*
>
> -John Tukey

5. 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.