Data Visualization and Effective Communication

Nicole A. Lazar

Department of Statistics
University of Georgia

Data Visualization: Essential for EDA and Beyond

EDA: Exploratory Data Analysis

Should be standard first step of any statistical analysis – simple tools such as boxplots, scatterplots, histograms, etc. as advocated by Tukey and others.

A large literature on this side of the equation, including principles of good statistical data visualization (Wainer, Tukey, Cleveland, Tufte . . .).

Simple examples bring the message home even to undergraduates or others with limited experience.

Example: Anscombe Data Sets

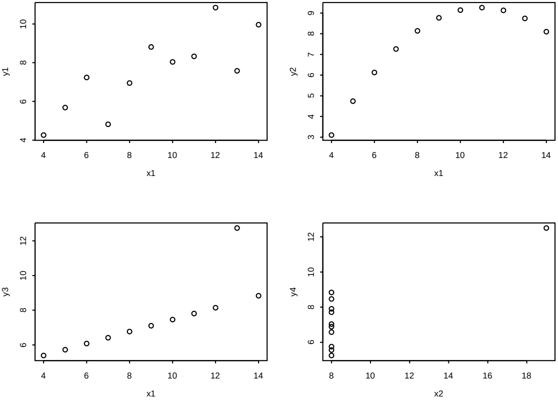
Data set	1-3	1	2	3	4	4
Variable	X	Y	Y	Y	X	Y
	10	8.04	9.14	7.46	8	6.58
	8	6.95	8.14	6.77	8	5.76
	13	7.58	8.74	12.74	8	7.71
	9	8.81	8.77	7.11	8	8.84
	11	8.33	9.26	7.81	8	8.47
	14	9.96	8.10	8.84	8	7.04
	6	7.24	6.13	6.08	8	5.25
	4	4.26	3.10	5.39	8	5.56
	12	10.84	9.13	8.15	8	7.91
	7	4.82	7.26	6.42	8	6.89
	5	5.68	4.74	5.73	19	12.50

Analysis of the Anscombe Data Sets

Basic summary statistics of all four data sets are the same:

- mean of X in both cases is 9;
- variance of X in both cases is 11;
- mean of Y in all four cases is 7.5;
- variance of Y in all four cases is 4.12;
- correlation between X and Y for all four data sets is 0.816;
- fitted regression line in all cases is $Y = 3 + 0.5X$.

The Anscombe Data Sets Plotted



The four Anscombe data sets.

ASA Guidelines on Learning Outcomes

At the Society level, what is advocated?

- Students should be able to **perform data analysis**: guidelines explicitly include graphical presentation of data (EDA).
- Students should be able to **communicate results**: guidelines include written and oral presentation skills, but no mention of data visualization.

Visualization is Part of Effective Statistical Communication

Gelman *et al.* (2002): use graphs not tables of data!

Tables of numbers can be (are) hard to process without careful study.

The message can often be conveyed more effectively with an appropriate plot.

This is true for presentation of research results, not just raw data.

Example 1: Someone Else

Table 1. *The number of genes that change their latent states from expressed to unexpressed and vice versa. The results for all 16 bra regions are shown.*

	Period 3-4	Period 4-5	Period 5-6	Period 6-7	Period 7-8	Period 8-9	Period 9-10	Period 10-11	Period 11-12	Period 12-13	Period 13-14	Period 14-15
MFC	0	10	92	515	359	132	114	90	45	9	0	0
OFC	0	20	88	525	354	135	117	89	45	7	0	0
VFC	0	16	72	524	356	134	114	91	47	7	0	0
DFC	0	15	76	522	354	136	115	89	48	7	0	0
STC	1	13	67	526	355	136	114	87	48	8	1	0
ITC	1	15	71	529	350	135	117	86	49	7	1	0
A1C	0	23	61	528	364	132	112	92	45	8	0	0
IPC	0	12	66	526	355	134	114	91	48	6	0	0
S1C	0	15	72	526	351	137	112	96	44	7	0	0
M1C	1	15	70	526	360	127	114	91	47	7	0	0
V1C	0	13	98	527	359	134	115	87	42	9	1	0
AMY	1	28	106	538	343	130	112	89	37	8	0	0
HIP	1	66	108	506	350	126	109	80	42	7	1	0
STR	1	34	72	511	347	115	114	79	45	9	0	0
MD	2	30	77	499	329	126	112	71	39	7	0	1
CBC	2	26	56	474	326	164	117	71	35	14	5	0

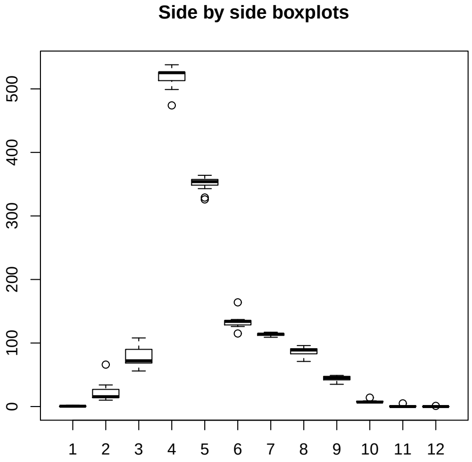
Example 1 Continued

Table shows numbers of genes differentially expressed in different brain regions over time.

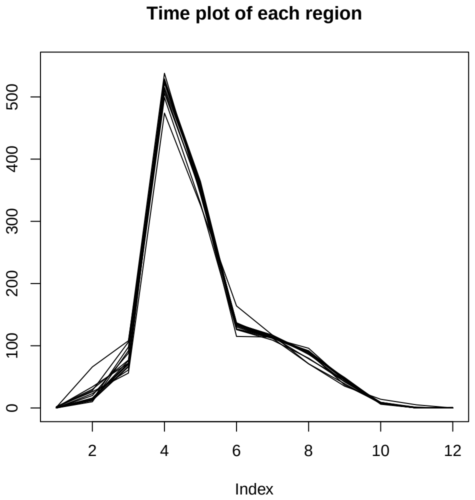
What is the take-home message of this table? Lots of numbers – are the specific values that important?

"Eyeballing" the patterns reveals commonalities. Why not a graphical presentation to make it clearer?

Example 1: Some Simple Graphical Presentations



Example 1: Some Simple Graphical Presentations



Example 2: One of My Former Students!

MSDR	Gau-100	Gau-150	Gau-200	Three-basis	Four-basis	Five-basis
61	0.1623	0.1622	0.5558	0.4671	0.7679	0.9611
62	0.2038	0.1905	0.4340	0.2114	0.5591	0.6899
63	0.2414	0.3821	0.6499	0.6545	0.9390	0.7945
64	0.3442	0.4448	0.6301	0.4059	0.6394	0.7436
65	0.2356	0.1928	0.4863	0.4960	0.5908	0.8987
66	0.5023	0.7017	0.9157	0.8081	0.9743	1.2199
67	0.3824	0.3794	0.6036	0.6648	0.8743	1.2336
68	0.2493	0.3406	0.5928	0.6634	0.5390	1.0485
69	0.8143	1.1373	1.4314	0.8004	0.9590	1.1643
70	0.1939	0.1482	0.4286	0.3198	0.4163	1.1748
71	0.1649	0.1678	0.5540	0.3882	0.6757	0.9063
72	0.2553	0.3342	0.5304	1.0922	0.5524	0.6870
73	0.4075	0.6141	0.9003	0.4205	0.8913	1.0452
74	0.2166	0.2648	0.5312	1.0820	0.6708	0.6594
75	0.3120	0.4904	0.8536	0.4647	0.9848	0.9745
76	0.1766	0.1160	0.3457	0.2793	0.3736	0.8522
77	0.8709	0.3477	0.4978	0.8225	1.2576	0.8071
78	0.2349	0.2540	0.4855	0.6369	0.7949	0.6010
79	0.6453	0.5982	0.9221	0.8624	0.9046	0.9034
80	0.2405	0.3073	0.5337	0.4678	0.6635	1.0207
81	0.3252	0.4411	0.7044	0.6559	1.1675	0.8254
82	0.4240	0.2898	0.3791	0.6646	0.8815	0.5478
83	0.1954	0.1831	0.6250	0.2075	0.5067	0.8635
84	0.4911	0.6099	0.7161	0.9085	1.2724	0.7338
85	0.2792	0.4200	0.6602	0.3675	1.1654	1.2136
86	0.2607	0.1908	0.3377	0.5706	0.4438	1.2555
87	0.2180	0.3365	0.5484	1.1003	0.6575	0.6170
88	0.3140	0.4890	0.7676	1.1869	0.5138	0.6278
Mean	0.3350	0.3762	0.6293	0.6311	0.7727	0.8954

Table 1: MSDR for different time points (61-88) under the Gaussian-type model approach and the nonparametric model approach. The difference between the two approaches is significant. See text for explanation.

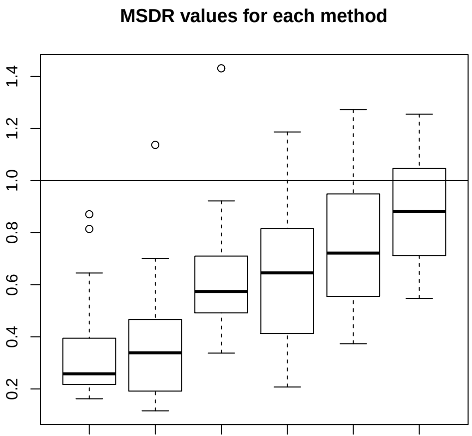
Example 2 Continued

Table shows measure of error for different fitting methods, at different time points in a neuroimaging data set.

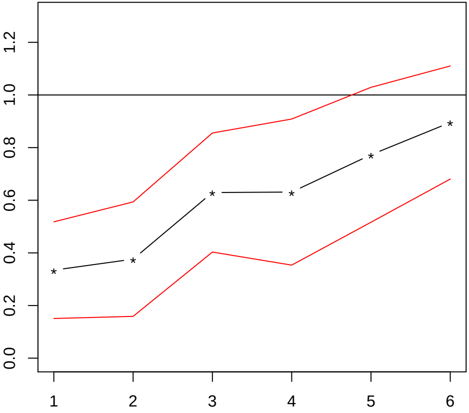
Values close to 1 indicate better performance – hard to pick out in the mass of numbers.

Is it necessary to look at each time point separately?

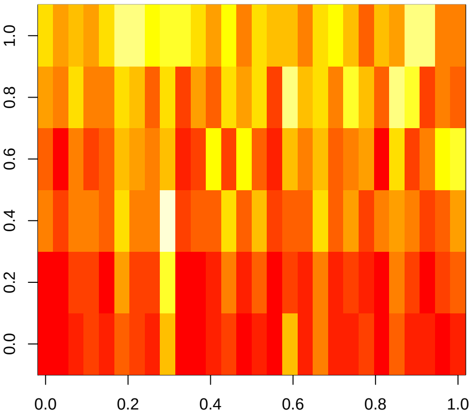
Example 2: Some Simple Graphical Presentations



Example 2: Some Simple Graphical Presentations



Example 2: Some Simple Graphical Presentations



Example 2 Continued

First three methods are parametric with increasing values of the parameter. Immediate conclusion: larger parameter value gives better fit.

Second three methods are nonparametric with increasing number of basis functions. Immediate conclusion: more basis functions give better fit.

Nonparametric methods give better fit overall than parametric methods.

Aside from some outliers, parametric methods are less variable in general.

Visualization Helps . . . But Plot Something Meaningful!

The "flip side" of the tables versus plots dilemma is a plot for the sake of a plot.

Or, more critically – a contentless plot.

Example: A Plot With No Content

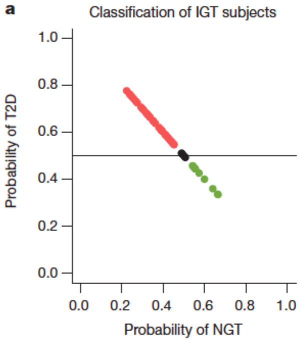


Figure 4 | Stratification of IGT women based on gut microbiota profiles. **a**, Use of the MGC model trained for discriminating NGT and T2D to classify IGT women ($n = 49$) as either NGT (green) or T2D (red).

Example: A Plot With No Content

What is plotted here?

Analysis of microbial communities in diabetic and healthy people leads to a prediction for which members of a third group will become diabetic.

Vertical axis gives probability of being Type 2 Diabetic; horizontal axis gives the probability of being healthy.

Probability of being healthy and probability of being Type 2 Diabetic add up to 1! So the graph **could only** be a straight line of slope -1.

Colors: red for individuals with probability greater than 0.5 of being Type 2 Diabetic; green for individuals with probability less than 0.5 of being Type 2 Diabetic.

Information to ink ratio of roughly zero . . .

Example: A Plot With No Content

This Figure appeared in *Nature*.

Big Data Can Exacerbate the Problem

With "Big Data" visualization can be particularly challenging – traditional graphical techniques may not (typically won't be) appropriate.

One implication: A need for statisticians to develop new analysis **and** visualization tools that are tailored to the application.

Another implication: Out of desperation confusing, contentless, or misleading graphical representations of data may be published.

Huge opportunity for us to make an impact here!

Example: Why Big Data Are Challenging

Functional magnetic resonance imaging (fMRI) data – data collected on the working of the human brain over time (on the scale of 10 minutes, often).

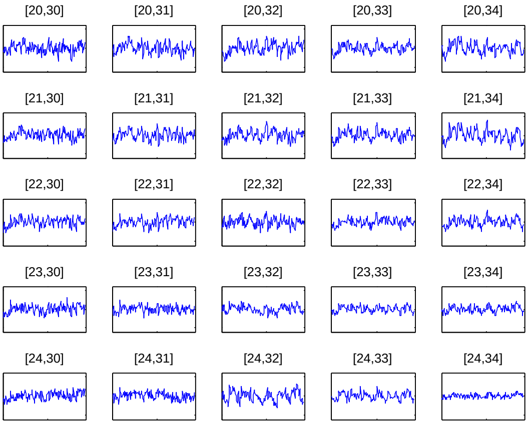
For a single individual:

- Multiple time points, usually on the order of several hundred.
- Multiple voxel locations, usually on the order of several hundreds of thousands.

Typical goal is to discover those voxels that are reacting to a particular task performed by the subject while in the MR scanner.

Example: A Small Piece of fMRI Data

Time courses for 25 voxels



Example: A Small Piece of fMRI Data

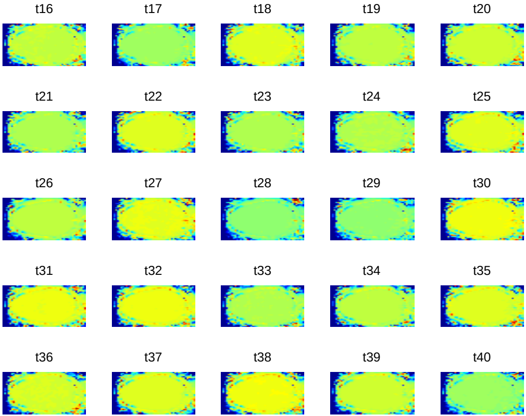
There are thousands of voxels – it's not feasible to visualize all the individual time courses **and** make sense of them.

The goal is to find those voxels with time courses that match (in some way) the design of the experiment – signal changes that correlate with changes in stimulus.

Needed: visualization techniques that rely on (sufficient) dimension reduction, principal components, clustering, etc.

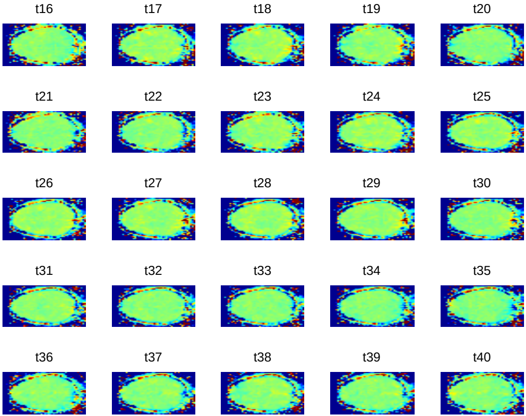
Example: A Small Piece of fMRI Data

Images for one slice of data, 25 time points, unscaled



Example: A Small Piece of fMRI Data

Images for one slice of data, 25 time points, scaled



Example: A Small Piece of fMRI Data

"Brain course" images are even harder to interpret, as from time point to time point it is difficult to see the changes.

Scaling makes a big difference here.

We are left with the difficulty of visualizing masses of data – and fMRI data are small(ish) by Big Data standards.

A Final Example: Everyone Is Doing It . . .

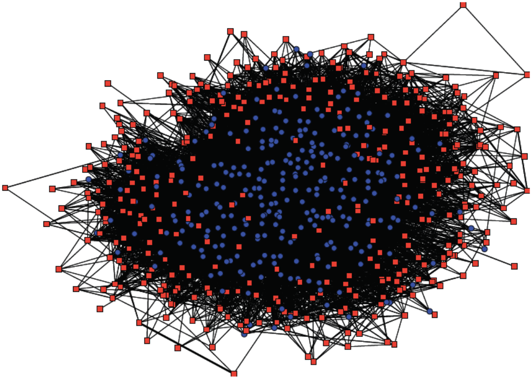


Fig. 2. Social network of whales sighted at least 20 times. Blue nodes are individuals observed lobtail feeding, red nodes are those never observed lobtail feeding. The network was laid out by spring-embedding using Netdraw (*25*) software.

A Final Example: Everyone Is Doing It . . .

Interactions of several hundred whales via more than 70,000 sightings. Analysis of the occurrence of “lobtail” tactic of fin-slapping shows cultural diffusion.

Data collected over three decades – what information can be mined from a massive data set such as this? And how to display?

Network analysis is very popular, and especially in the Big Data setting. But what does the network graph mean for **these** data?

Conclusions

Visualization is an important part of the statistician's toolbox, both for exploratory data analysis and presentation of our own research results.

We do a pretty good job at introducing the former, but even now, are not as effective in emphasizing the latter (to our students, in our own practice ...).

Big Data poses new and exciting challenges for data visualization and communication of large, complicated structures.

Plenty of opportunity for us as a community to make contributions in this area.