Lecture 3: Exploratory Data Analysis (EDA)

Olexandr Isayev

Department of Chemistry, CMU

olexandr@cmu.edu

Outline

- Part 1: Data, data types, formats, basic analysis (last week)
- Part 2: Data Visualization (Today)
- Part 3: Dimensionality reduction (Next Tuesday)

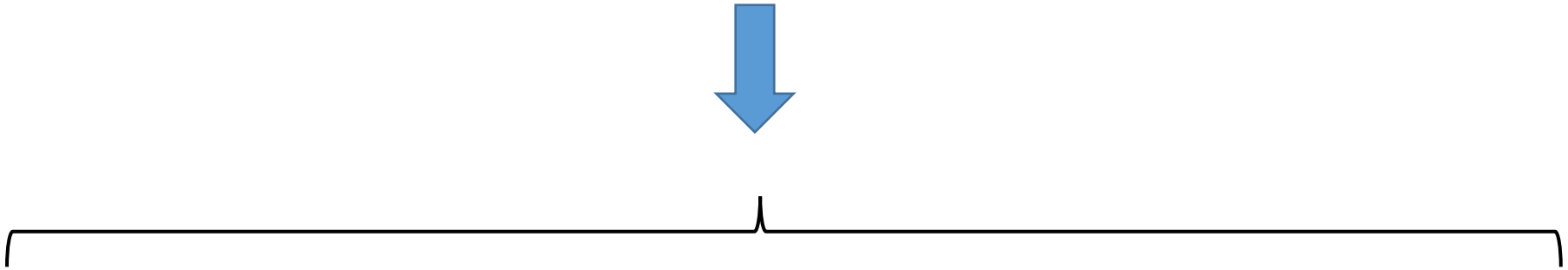
Lab/Home assignment 1: EDA

We provided you with an obfuscated scientific dataset.

- Each row is an observation.
- The first column, titled “experimental_proprty” is your target property of interest.
- All other columns encode features.
- There are no observation IDs.

Features

Target variable



experimental_proprty	MS_enc	nHetero	nX	C2SP3	MPC5	ATS0i	NaasN	SsCH3	nCl	nFAHRing	SlogP_VSA	AATS3d	ATSC6are	nHBDon	SddssS	ATS6m	nAcid	PEOE_VSA	SsssCH
3.54	PPENPINE	5	1	0	83	7305.483	1	2.105782	1	0	48.53094	3.127273	-0.05981	0	0	4833.659	0	43.93672	0
-1.18	PPENPINE	11	0	3	114	9532.634	0	4.605907	0	1	41.29369	3.477273	-0.77856	2	-4.06028	12049.8	1	12.10821	-0.27577
3.69	PPENPINE	5	1	2	69	5845.024	0	1.422763	1	1	35.7119	3.4	-2.08499	0	0	6321.555	0	22.98929	-0.44745
3.37	HTXPTDW	9	1	3	105	7388.625	0	0	1	1	36.3982	3.309735	0.727528	4	0	7928.349	0	17.49443	-1.82906
3.1	PPENPINE	7	0	2	68	9082.561	1	9.734779	0	0	30.33184	2.816	1.267383	2	0	7691.124	0	18.88348	-0.79571
3.14	PPENPINE	2	0	2	81	6942.644	0	0	0	1	54.5973	2.932773	-0.1349	1	0	3797.86	0	18.02611	0.336659
-0.72	PPENPINE	11	0	5	103	10814.72	0	2.704557	0	0	41.29369	3.089744	-2.216	3	-4.05219	11716.26	1	37.80884	0
0.34	PPENPINE	9	2	3	84	9504.644	0	3.704763	2	0	54.5973	3.068182	0.05155	3	0	9139.272	1	25.58324	-1.15599
3.05	GOMNOO	5	0	0	85	6168.237	0	4.648601	0	0	51.87305	3.932432	-4.63798	0	0	4546.707	0	12.13273	0
2.25	PPENPINE	5	0	0	51	3200.04	0	0	0	0	28.53388	4.162162	-1.15247	1	0	2412.855	0	10.94351	0
1.51	PPENPINE	6	0	2	91	8102.957	0	1.175539	0	1	53.42642	2.947761	2.369929	1	-3.21631	5430.293	0	24.28188	0.266216
2.61	PPENPINE	9	1	1	74	7292.579	0	5.48351	0	0	24.39594	3.261364	-0.22364	3	0	6420.612	0	11.7603	-0.2915
-0.08	PPENPINE	2	0	4	9	3191.552	0	0	0	0	0	2.019231	0	1	0	0	0	12.96558	0
1.95	PPENPINE	5	0	4	69	7092.07	0	2.136284	0	0	47.88196	3.028571	-1.91092	0	0	6270.992	0	25.13529	0.147899
1.34	PPENPINE	8	0	5	104	9729.403	1	4.029277	0	0	65.68963	3.013072	1.5899	2	-3.9212	10782.74	0	6.041841	0.446712
3.2	PPENPINE	4	0	0	59	4212.764	0	0	0	0	42.46457	3.381818	0.090472	2	0	2487.898	0	11.25084	0
1.6	PPENPINE	7	0	4	81	9241.69	0	1.65129	0	0	24.26547	2.90604	0.041676	1	0	6355.507	0	43.70222	-0.57482
3.77	PPENPINE	6	0	1	73	9290.584	0	10.60184	0	0	42.46457	3.333333	2.502037	0	0	7070.008	0	0	0
3.15	PPENPINE	9	1	2	122	7979.715	1	1.587723	1	0	53.32547	3.538462	1.831988	1	0	9463.415	0	29.05566	-0.33894
0.32	PPENPINE	10	0	3	129	10483.59	1	1.713774	0	1	47.25911	3.202454	1.084735	2	0	8742.512	0	45.39703	0.127067
2.92	PPENPINE	6	0	6	74	10543.22	0	4.488685	0	0	42.46457	2.53125	-0.98503	3	0	5806.559	0	12.14581	0
1.92	PPENPINE	10	1	0	84	7830.57	0	0	1	0	52.94809	3.205607	-1.99871	2	0	6795.175	0	35.05037	0
3.17	RCRMUKS	7	0	3	97	9411.366	1	3.363928	0	0	64.31685	3.06383	-0.21484	1	0	7199.475	0	24.84981	0.051418
2.17	PPENPINE	7	2	5	93	9535.912	0	0	2	0	42.46457	2.82716	0.695543	1	0	7284.99	1	44.47754	0.905831

Perform EDA analysis

- Load data, prepare for analysis, process if necessary
- Analyze types of data
- Find and process missing and erroneous features
- Find outliers (if any)
- Find highly correlated variables (if any).
- Find if the target variable is correlated with any features.
- Use PCA to plot data in 2D and color code by the target property. Do you see any patterns?
- Prepare a short write-up describing your processing technics and choices above. (use Jupyter to insert text area if you like)

Bonus Questions

- Use any non-linear dimensionality reduction method. Plot data in 2D and color code by the target property. Compare observed picture with PCA.

- Surprise me! Uncover hidden patterns and find non-trivial relationships in the data

Due Date September 18 at 11:59

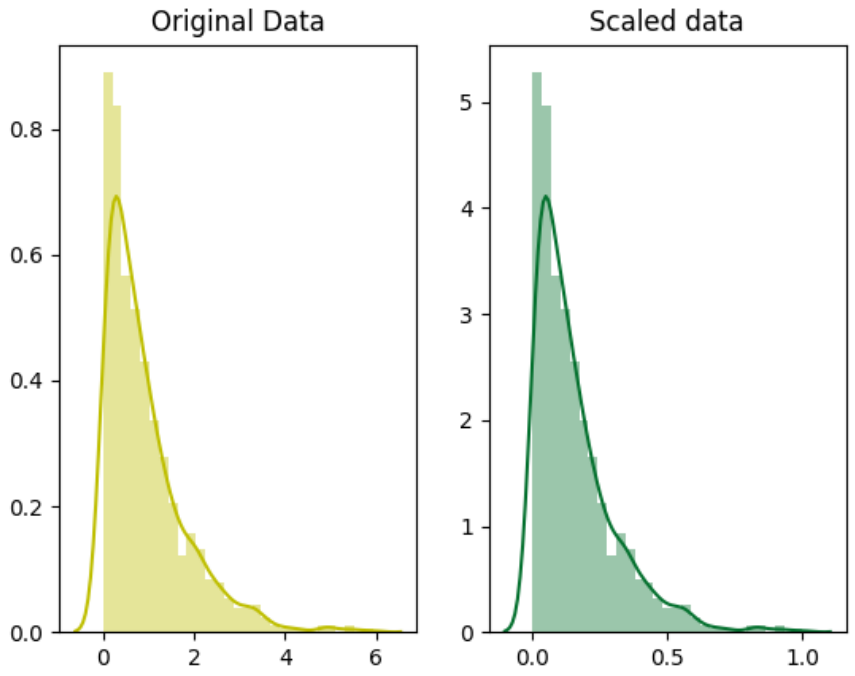
- Upload your Jupyter Notebook solutions to Canvas
- I will evaluate them quickly
- Thursday September 22 – in class discussions of your solutions

Basic Data Types in Statistics

Scaling

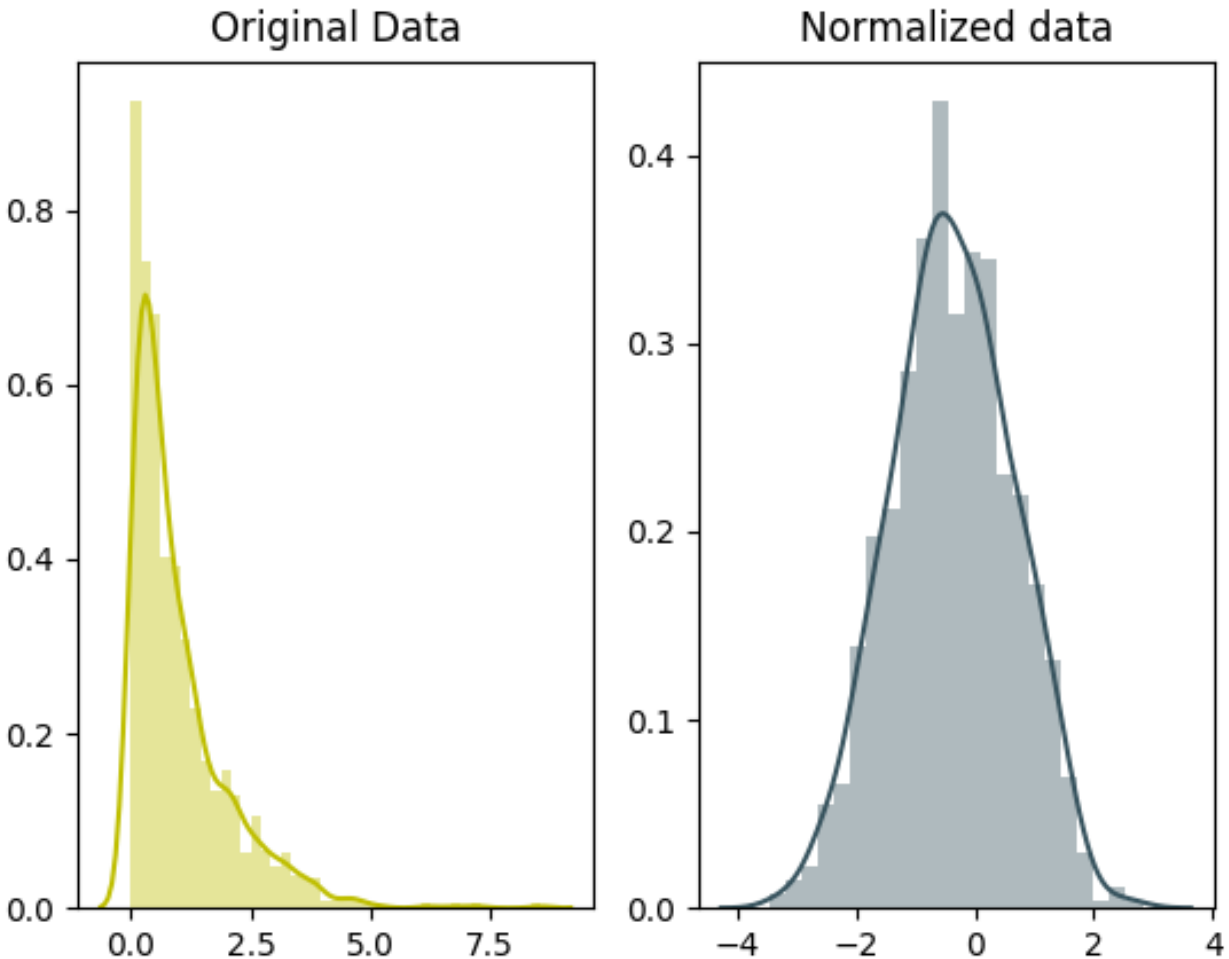
In scaling *(also called **min-max scaling**)*, you transform the data such that the features are within a specific range e.g. [0, 1].

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$



Good practice: always scale continuous data

Standardization



$$x' = \frac{x - x_{mean}}{\sigma}$$

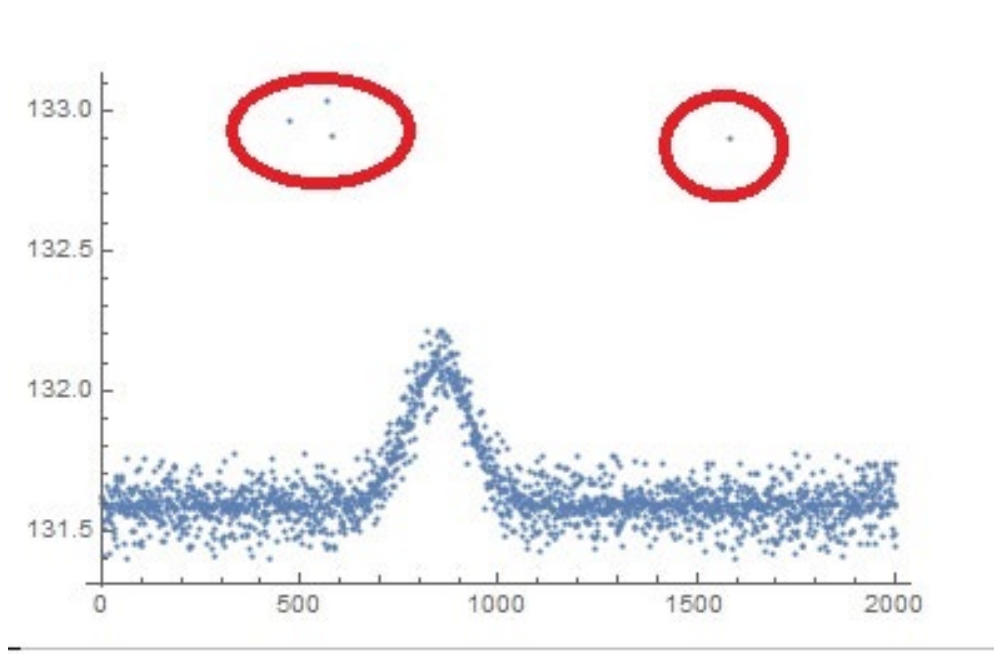
Outliers in Data

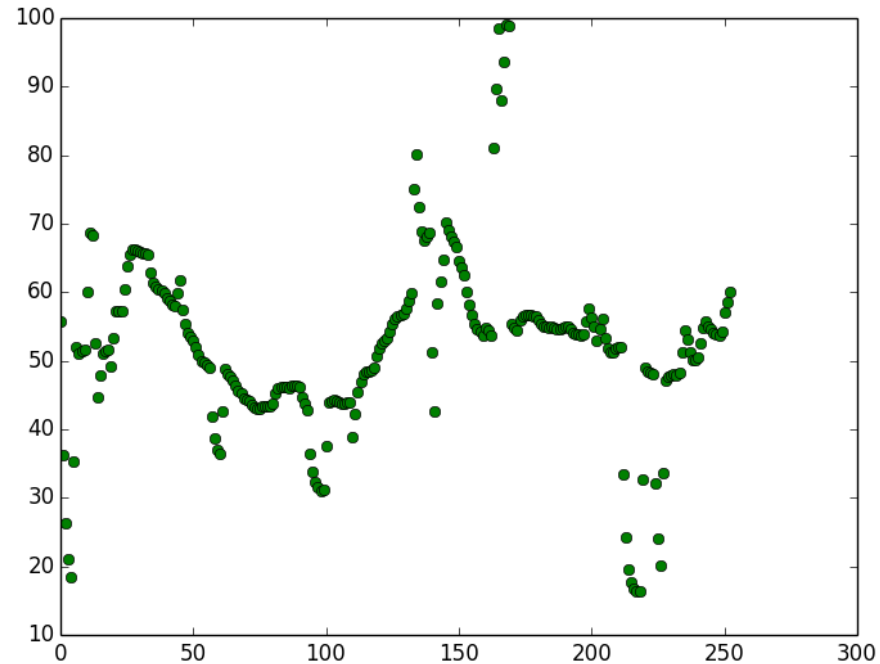
A measurement error or data entry error, correct the error if possible. If you can't fix it, remove that observation because you know it's incorrect.

Not a part of the population you are studying (i.e., unusual properties or conditions), you can legitimately remove the outlier.

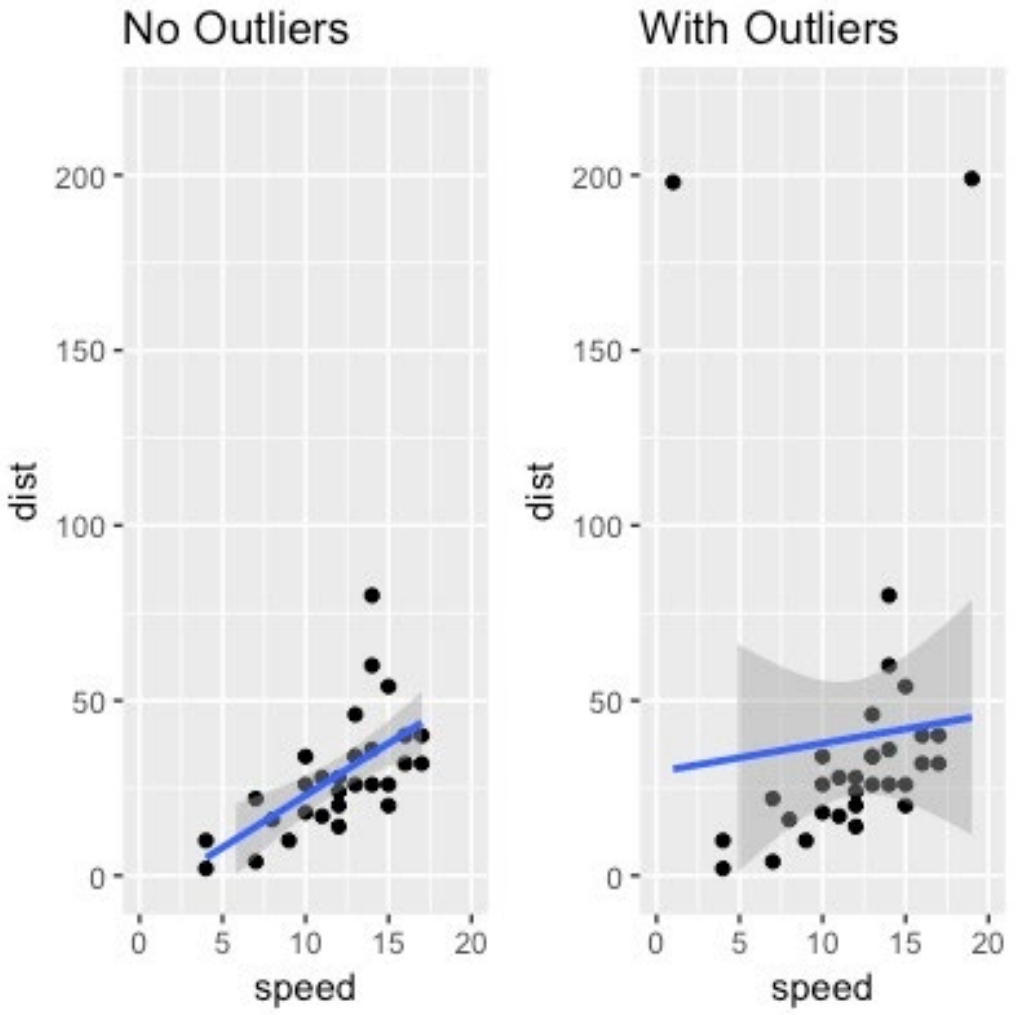
Sometimes it's best to keep outliers in your data:

- A natural part of the population you are studying, you should not remove it.
- They can capture valuable information that is part of your study area.


133.0
132.5
132.0
131.5
0
500
1000
1500
2000


100
90
80
70
60
50
40
30
20
10
0
50
100
150
200
250
300

Outliers Affect Statistics



- Basics of visualization
- Data types and visualization types
- Software plotting libraries: **matplotlib** and **seaborn**

Two types of visualization

Data exploration visualization: figuring out what is true

Data presentation visualization: convincing other people it is true

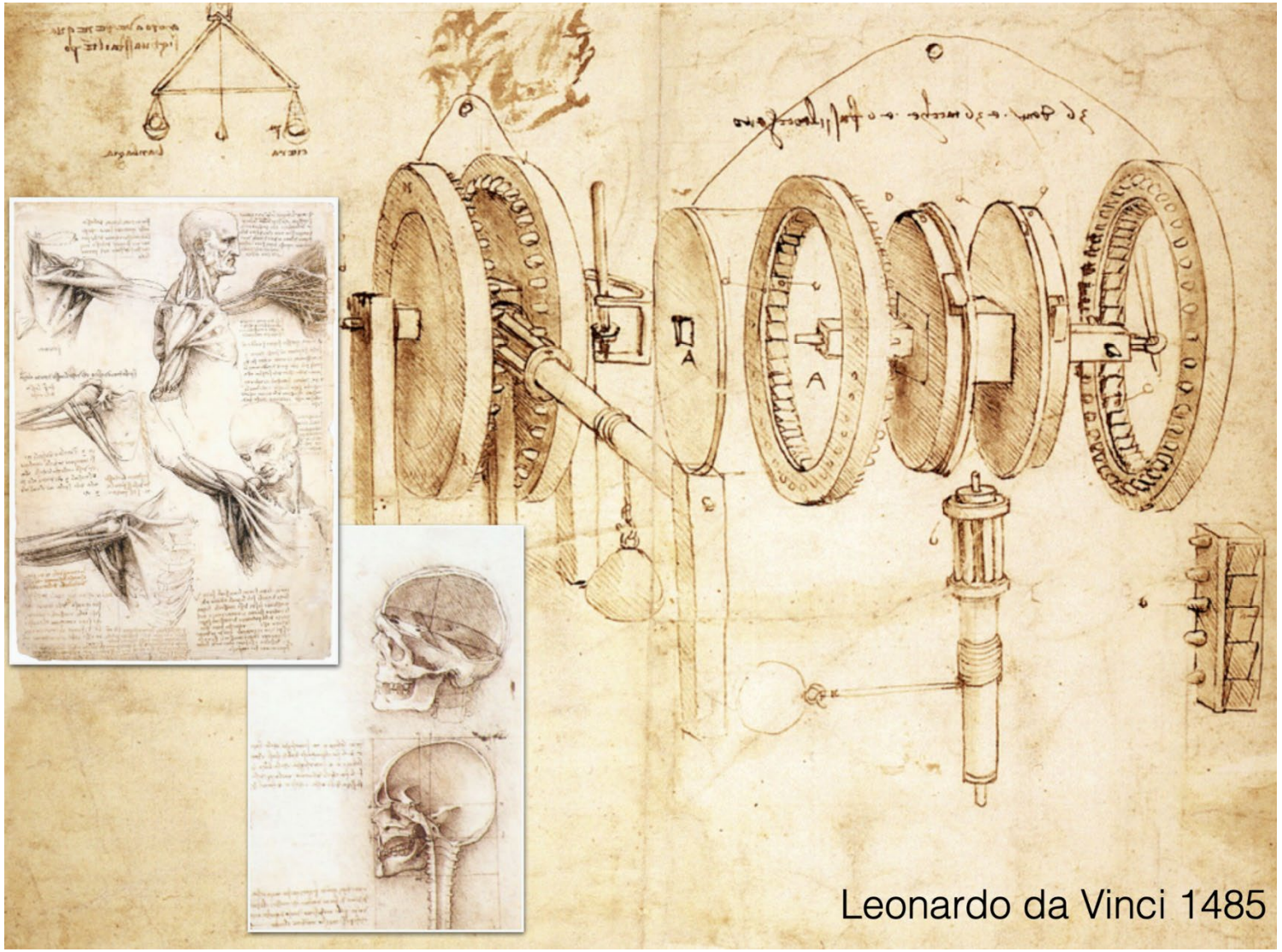
Importance of visualization

Before you run any analysis, build any machine learning system, etc, always

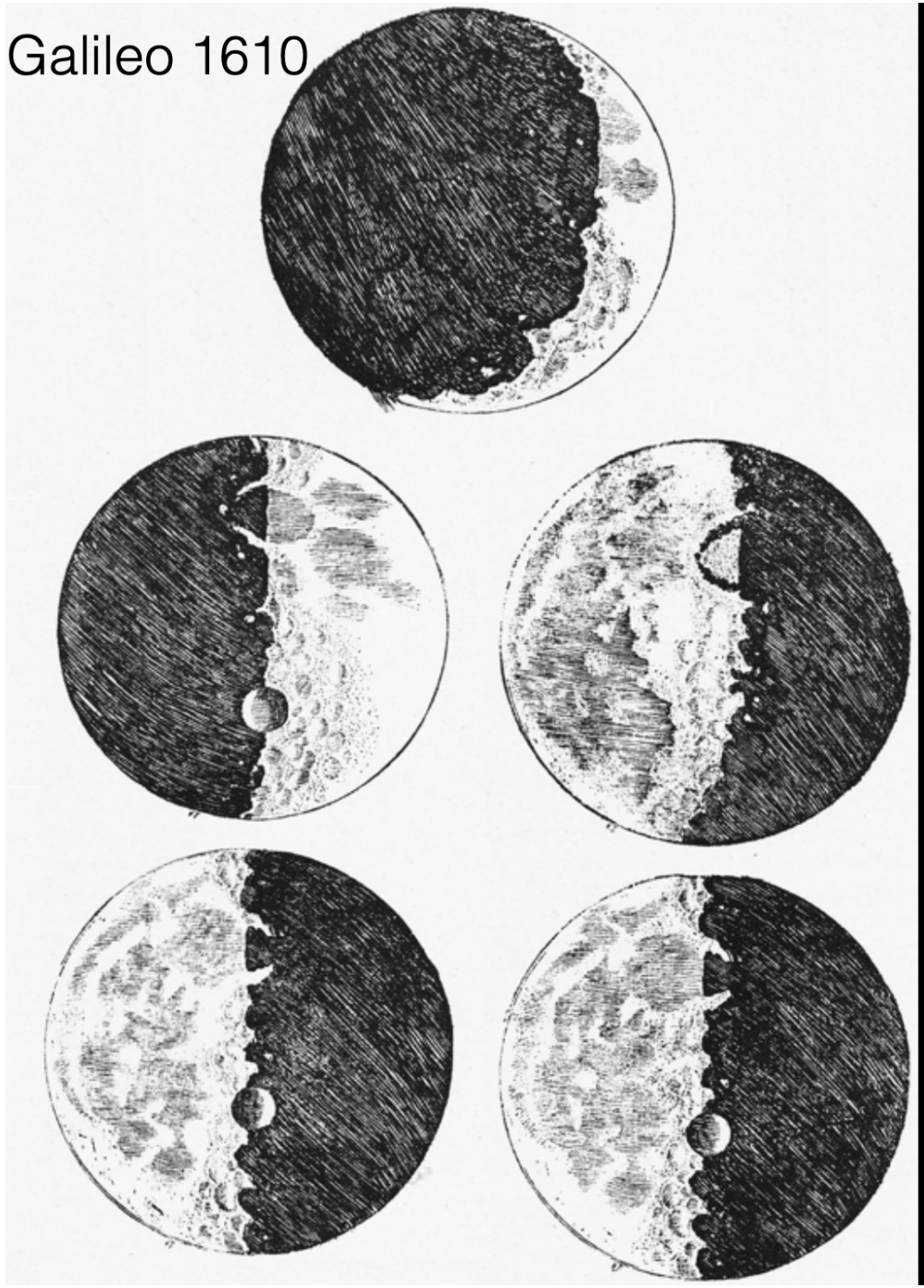
- visualize your data

- If you can’t identify a *trend* or make a prediction for your dataset, neither will an automated algorithm

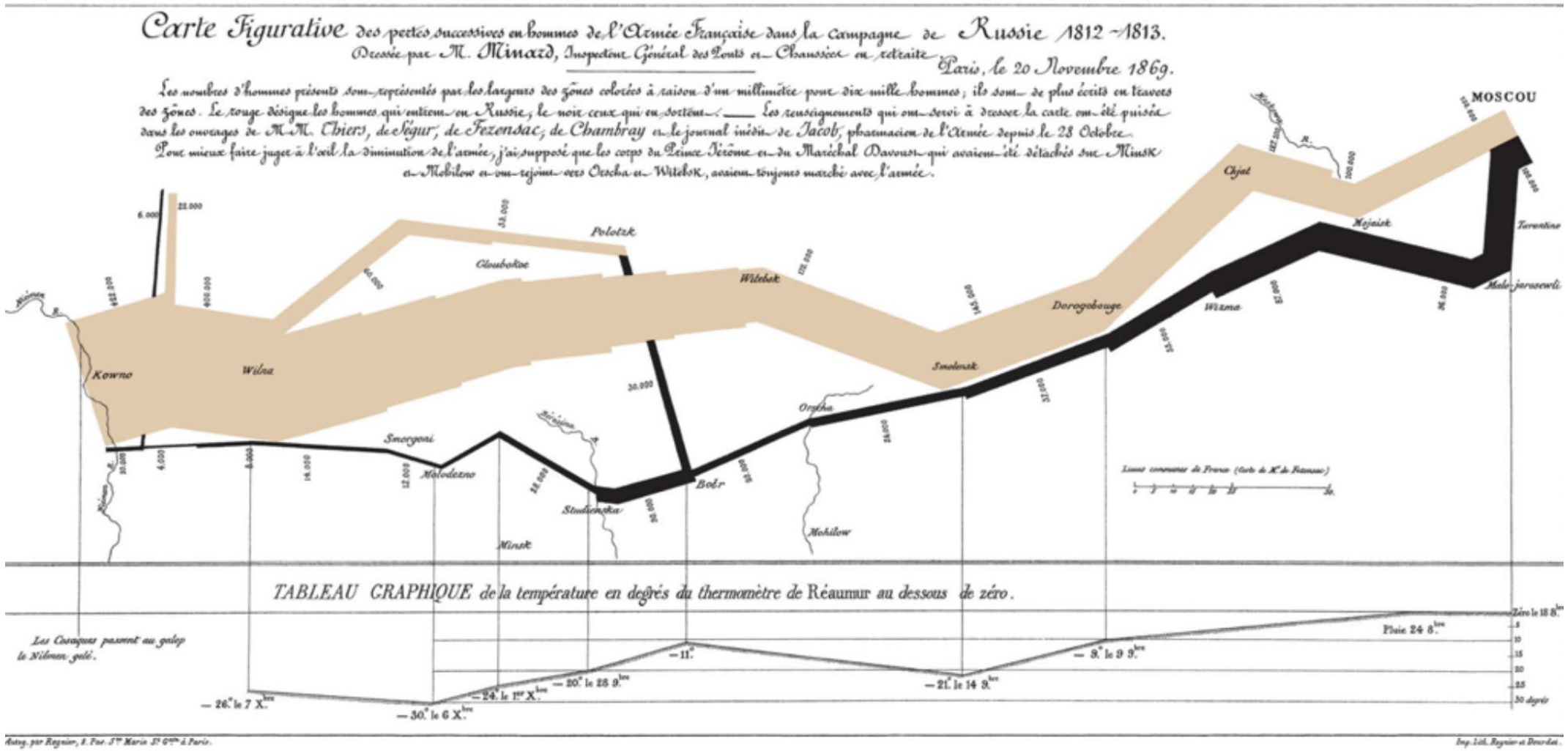
Good and Bad Visualizations



Galileo 1610



Joseph Minard 1861



It displays six types of data in two dimensions: the number of Napoleon's troops; the distance traveled; temperature; latitude and longitude; the direction of travel; and location relative to specific dates without making mention of Napoleon.

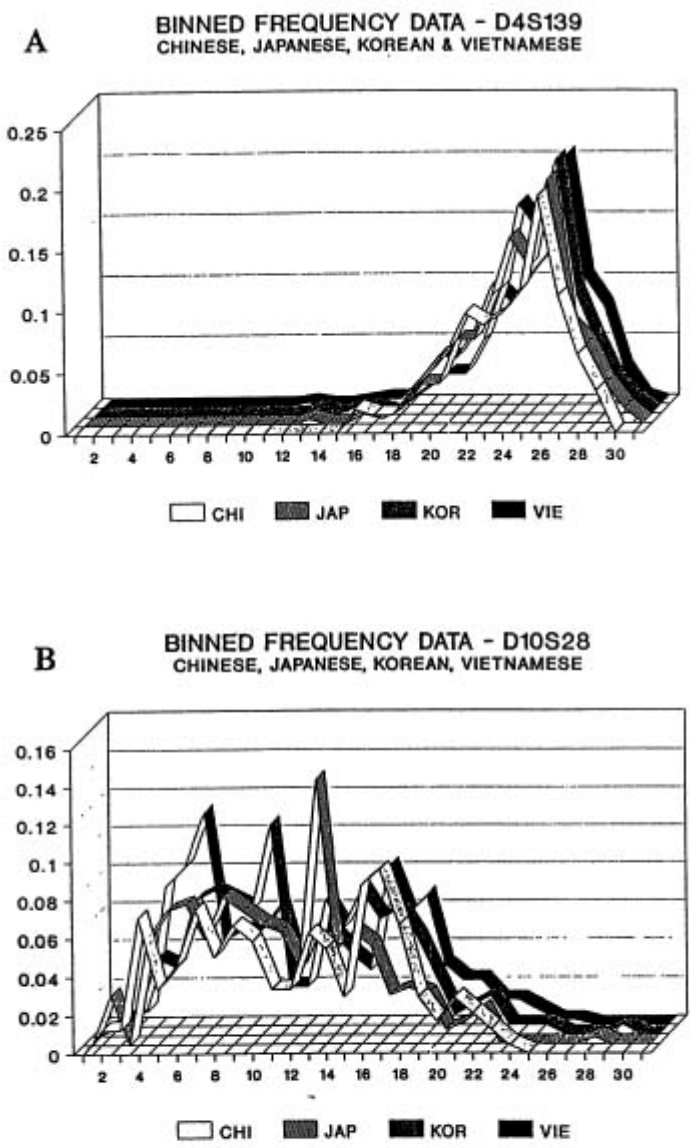
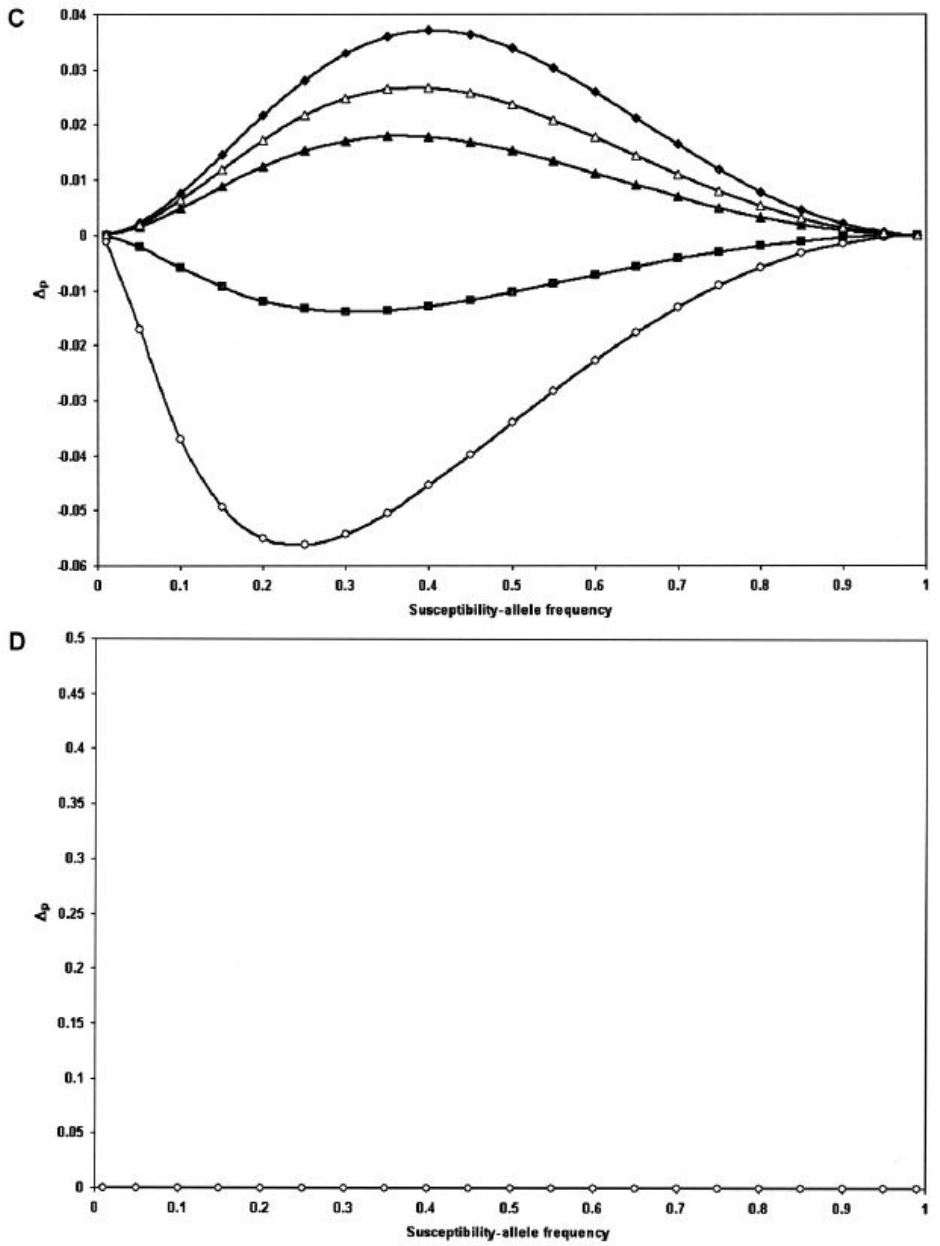


FIG. 4. *Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the* 30 *bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are* 103, 125, 93 *and* 215 *for* D4S139 *and* 120, 137, 100 *and* 193 *for* D10S28. *The horizontal axis is the bin number; bins are not of equal length.*

Roeder K (1994) DNA fingerprinting: A review of the controversy (with discussion). *Statistical Science* 9:222-278, Figure 4



Wittke-Thompson JK, Pluzhnikov A, Cox NJ (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *American Journal of Human Genetics* 76:967-986, Figure 1

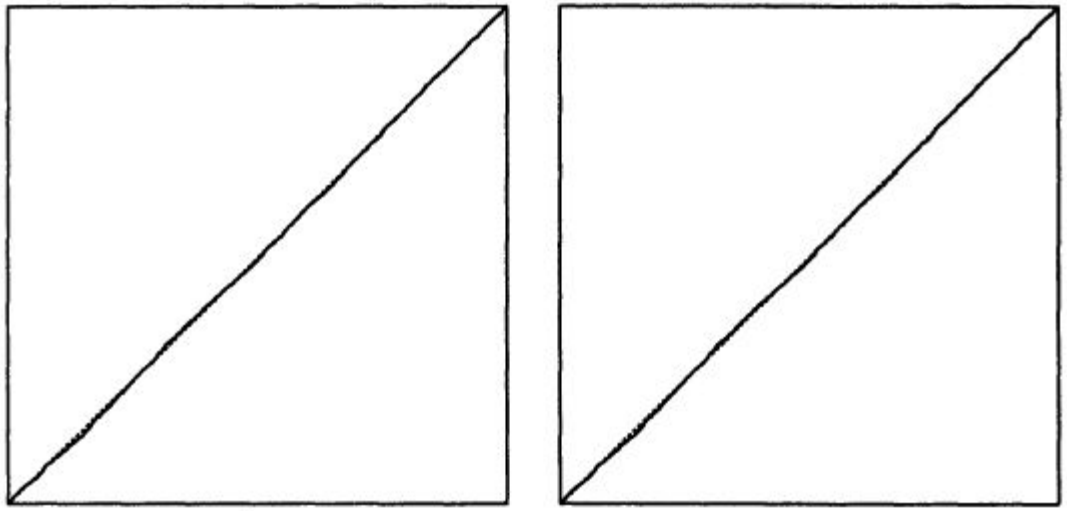


Figure 1. SRQ Plots of T_i/T_n (Vertical Axes) Against i/n (Horizontal Axes) for the Gibbs Sampler (a) and an Alternating Gibbs/Independence Sampler (b) for the Pump Failure Data Based on Runs of Length 5,000. Lines through the origin with unit slope are shown dashed; axis ranges are from 0 to 1 for all axes.

Mykland P, Tierney L, Yu B (1995) Regeneration in Markov chain samplers. *Journal of the American Statistical Association* 90:233-241, Figure 1

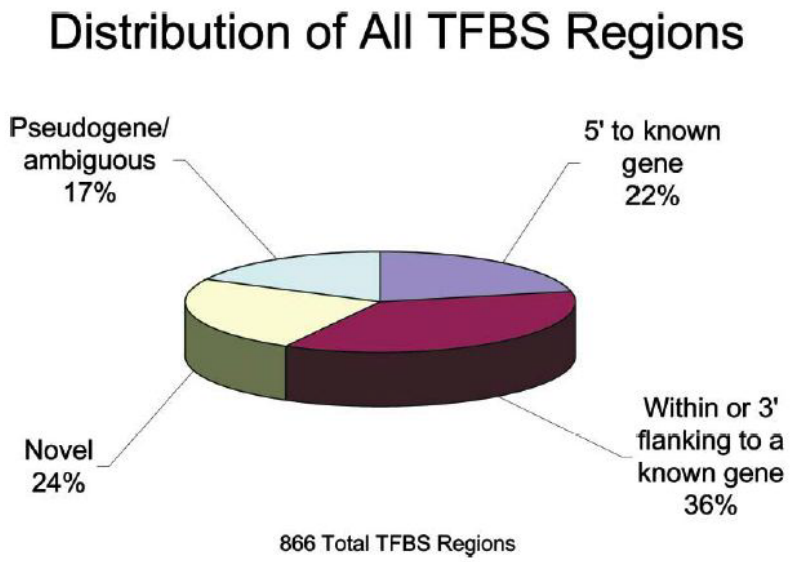


Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5′ most exon of a gene, within 5 kb of the 3′ terminal exon, or within a gene, novel or outside of any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).

Cawley S, et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116:499-509, Figure 1

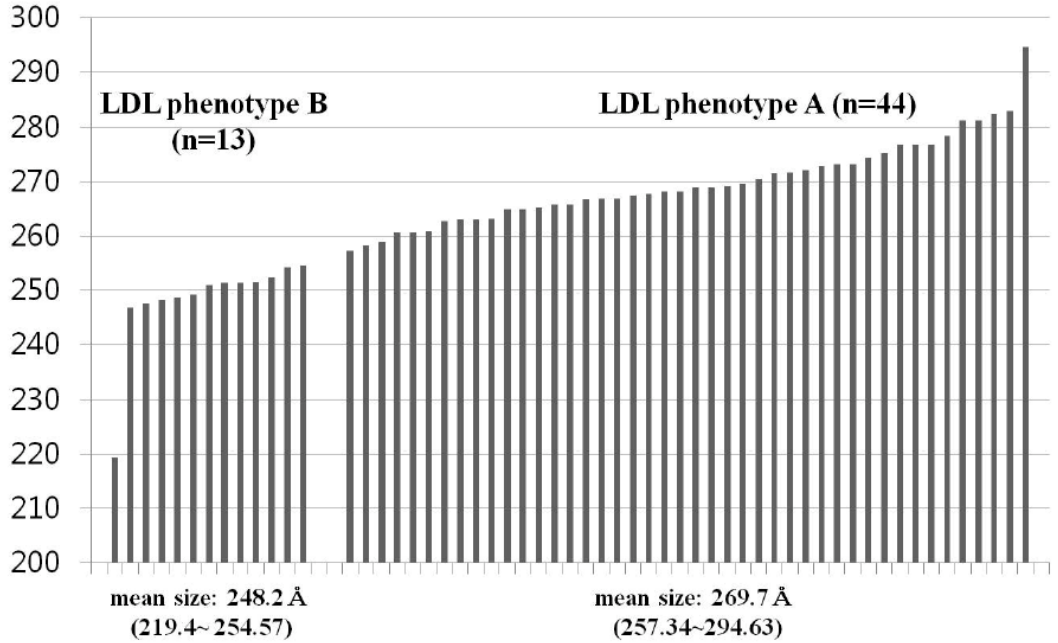
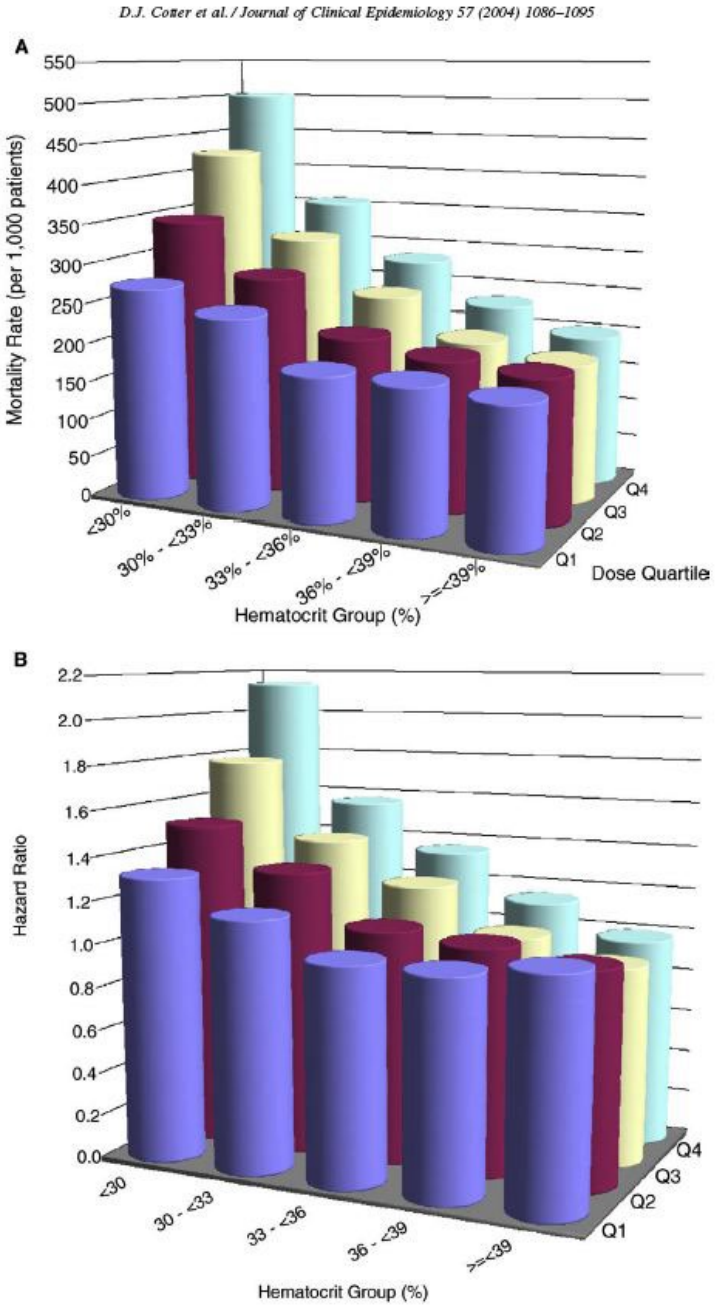
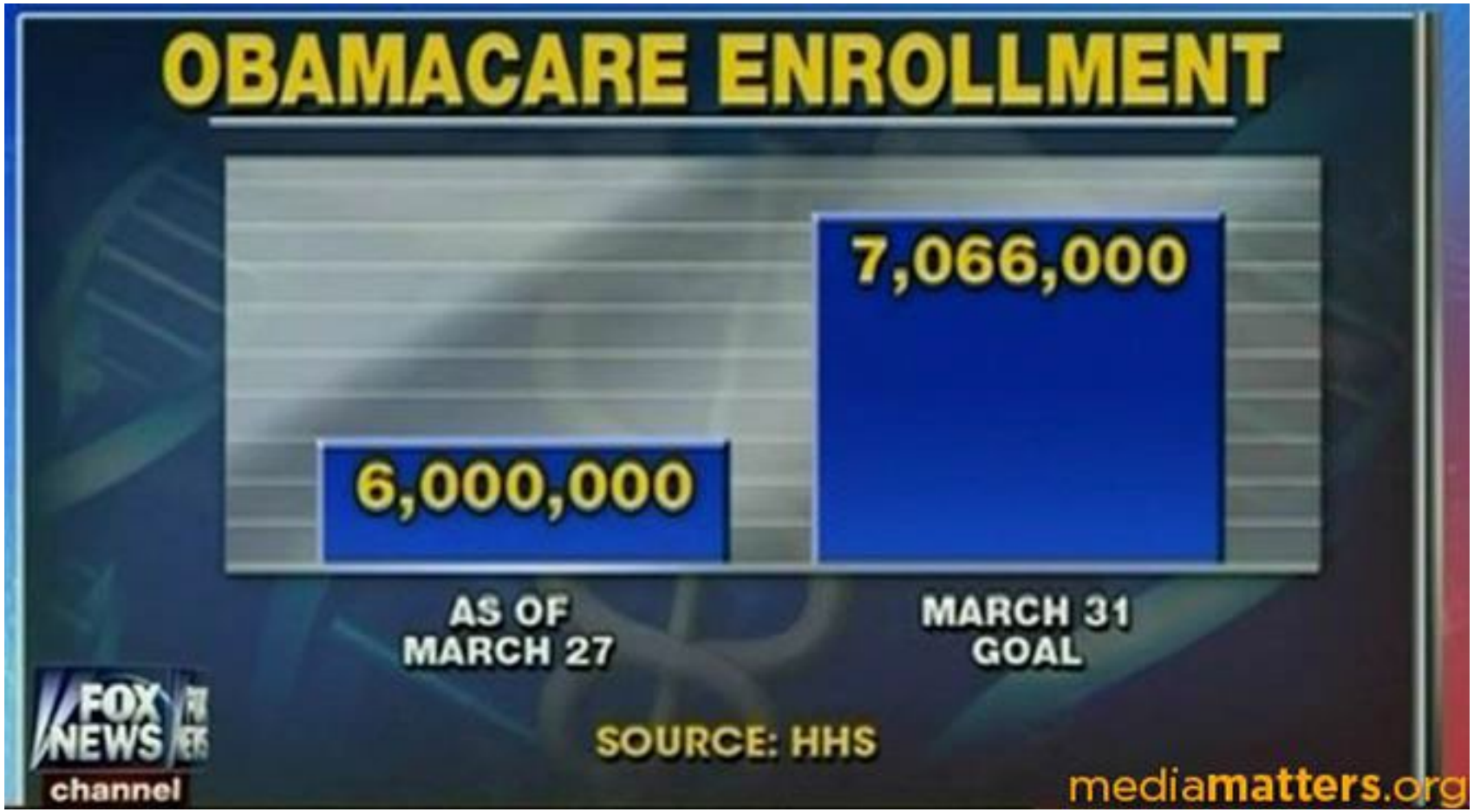


Fig. 1. Distribution of low-density lipoprotein (LDL) particle size in all study subjects (LDL phenotypes A and B). *LDL phenotype A group* (mean size: 269.7 Å, n = 44), subjects with buoyant-mode profiles [peak LDL particle diameter $\geq$ 264 Å] including intermediate LDL subclass pattern [256 Å $\leq$ peak LDL particle diameter $\leq$ 263 Å]; *LDL phenotype B group* (mean size: 248.2 Å, n = 13), subjects with dense-mode profiles [peak LDL particle diameter $\leq$ 255 Å]

Kim OY, et al. (2012) Higher levels of serum triglyceride and dietary carbohydrate intake are associated with smaller LDL particle size in healthy Korean women. *Nutrition Research and Practice* 6:120-125, Figure 1



Cotter DJ, et al. (2004) Hematocrit was not validated as a surrogate endpoint for survival amoung epoetin-treated hemodialysis patients. *Journal of Clinical Epidemiology* 57:1086-1095, Figure 2

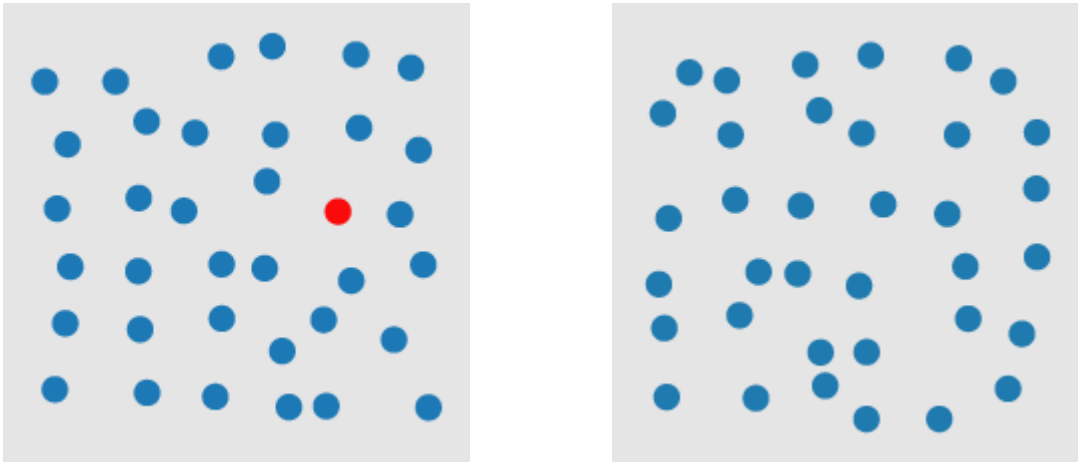

OBAMACARE ENROLLMENT
6,000,000
7,066,000
AS OF MARCH 27
MARCH 31 GOAL
SOURCE: HHS
FOX NEWS channel
mediamatters.org

Why Use Visualization?

- Vision is the most powerful communication channel humans' posses.

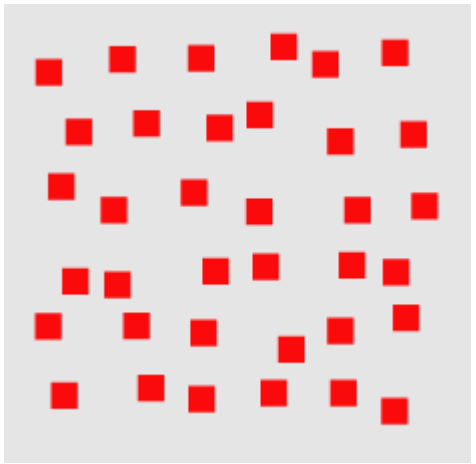
- For instance, We can detect information faster than our eye can move

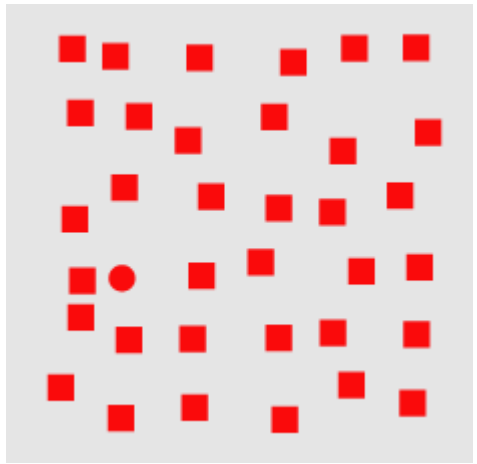
Preattentive Processing



Preattentive features can be detected faster than eye movement (200 msec).

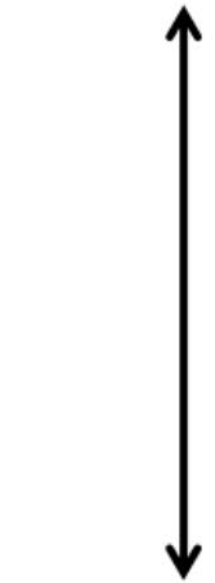
Compare to this…





Ranking of Visual Variables

More Accurate



Less Accurate

Quantitative	Ordinal	Nominal
Position	Position	Position
Length	Density	Hue
Angle	Saturation	Density
Slope	Hue	Saturation
Area	Length	Shape
Density	Angle	Length
Saturation	Slope	Angle
Hue	Area	Slope
Shape	Shape	Area

Cleveland, William S., and Robert McGill. "Graphical perception: Theory, experimentation, and application to the development of graphical methods." *Journal of the American Statistical Association* 79.387 (1984): 531-554.

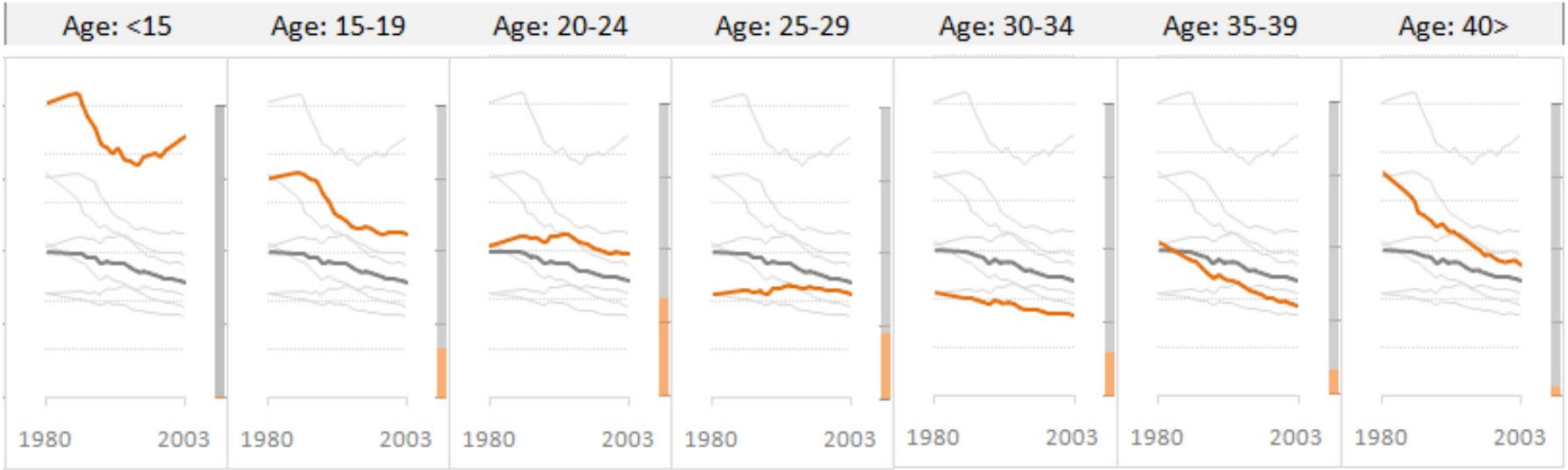
Humans are not very good at detecting patterns from numbers.

	1980	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
Less than 15 years old	607	624	605	578	553	523	515	502	511	492	488	479	493	498	504	497	512	519	529	537
15 to 19 years old	451	462	457	449	444	418	403	379	370	364	353	347	350	346	341	337	339	341	339	337
20 to 24 years old	310	328	328	327	327	318	328	330	333	334	326	317	314	307	301	297	296	298	296	293
25 to 29 years old	213	219	219	216	218	213	224	224	228	230	227	224	228	226	224	221	220	219	215	211
30 to 34 years old	213	203	201	197	194	189	196	192	192	189	183	179	178	176	174	171	169	171	169	167
35 to 39 years old	317	280	277	265	254	244	249	241	239	234	226	219	215	208	203	200	195	195	190	186
40 years old and over	461	409	381	374	361	350	354	339	338	329	320	309	301	291	290	283	276	276	278	268

Which group has the highest/lowest rates? When?

Which group has an increasing/decreasing temporal trend?

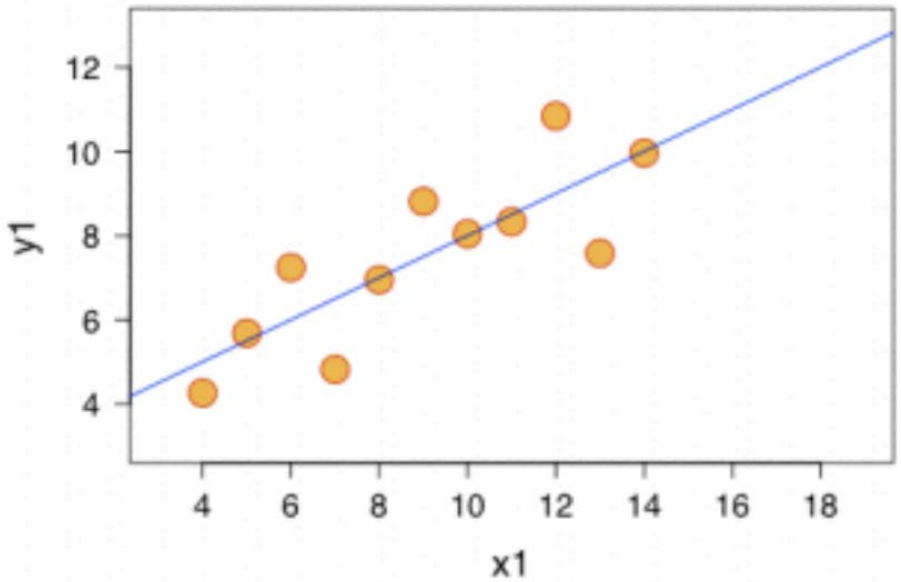
Which group has a faster/slower rate of change?

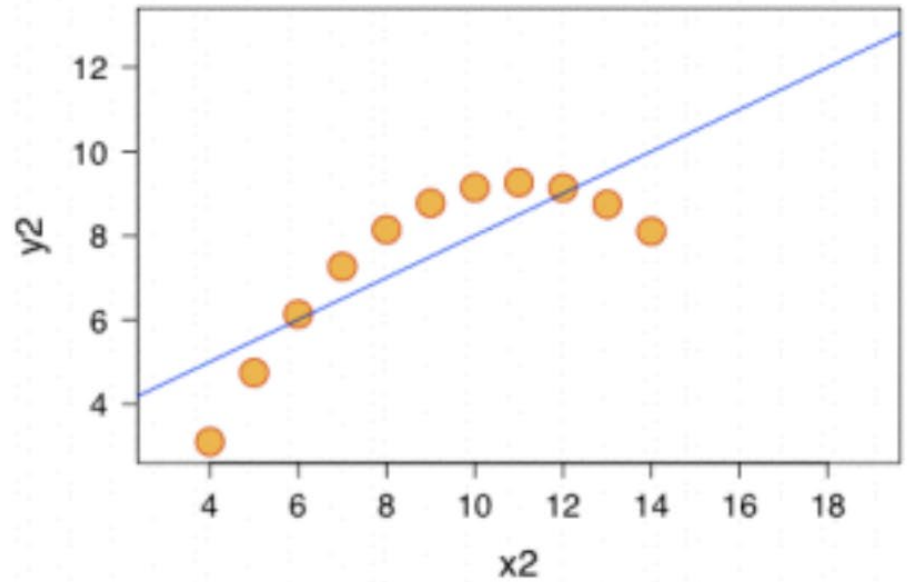

Age: <15
Age: 15-19
Age: 20-24
Age: 25-29
Age: 30-34
Age: 35-39
Age: 40>
1980
2003

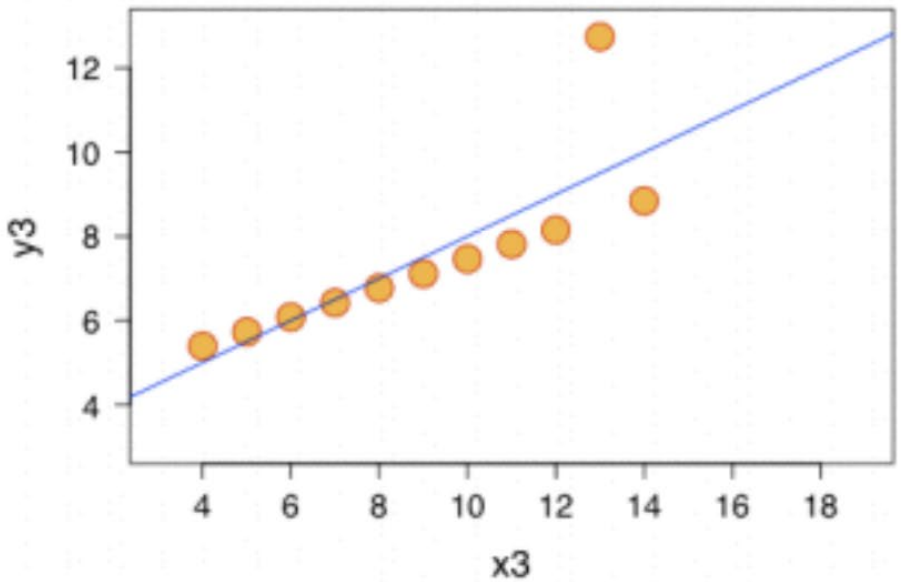
Summary statistics can hide important information.

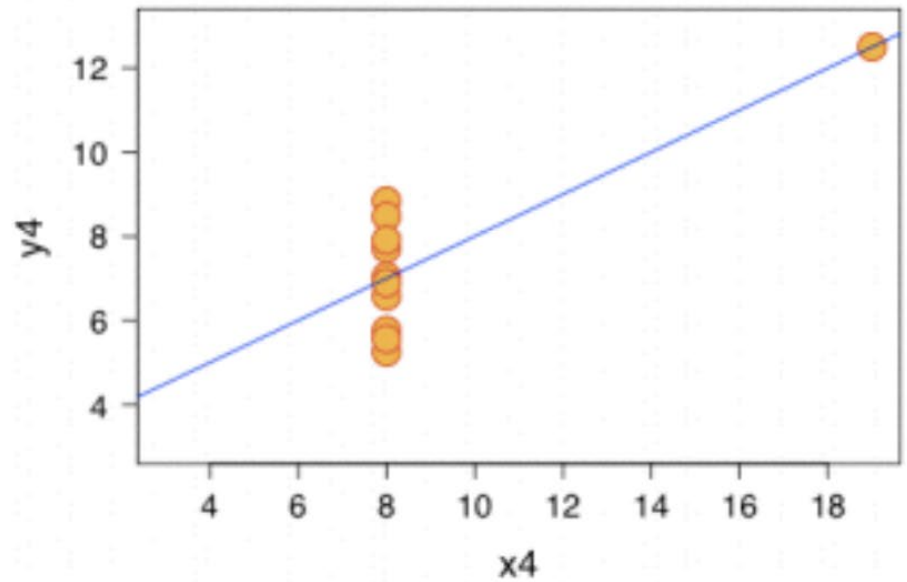
Anscombe's Quartet: Raw Data

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
corr.	0.816		0.816		0.816		0.816	


y1
x1
4
6
8
10
12
14
16
18
4
6
8
10
12


y2
x2
4
6
8
10
12
14
16
18
4
6
8
10
12


y3
x3
4
6
8
10
12
14
16
18
4
6
8
10
12


y4
x4
4
6
8
10
12
14
16
18
4
6
8
10
12

These six datasets have the same summary stats.

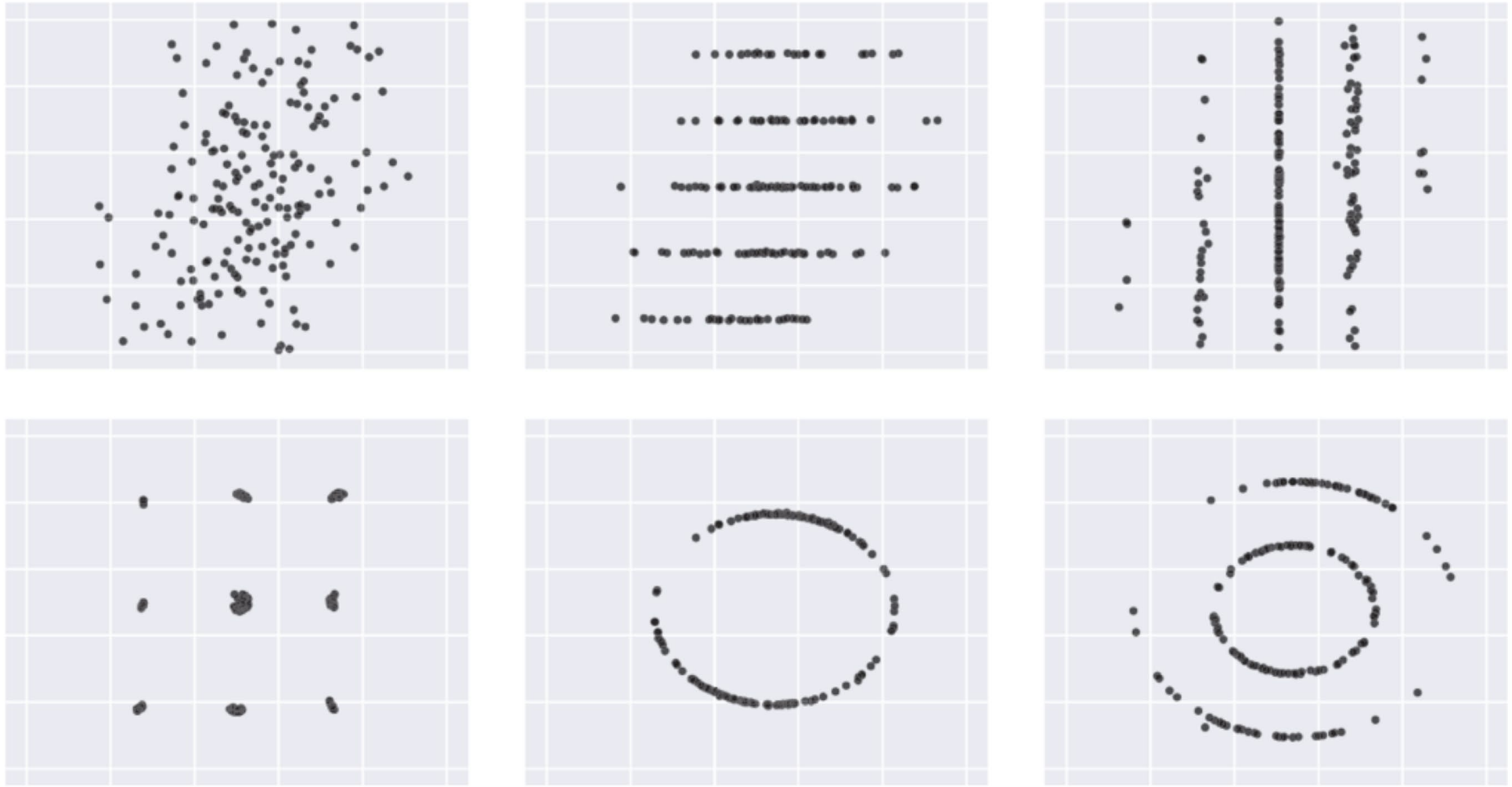
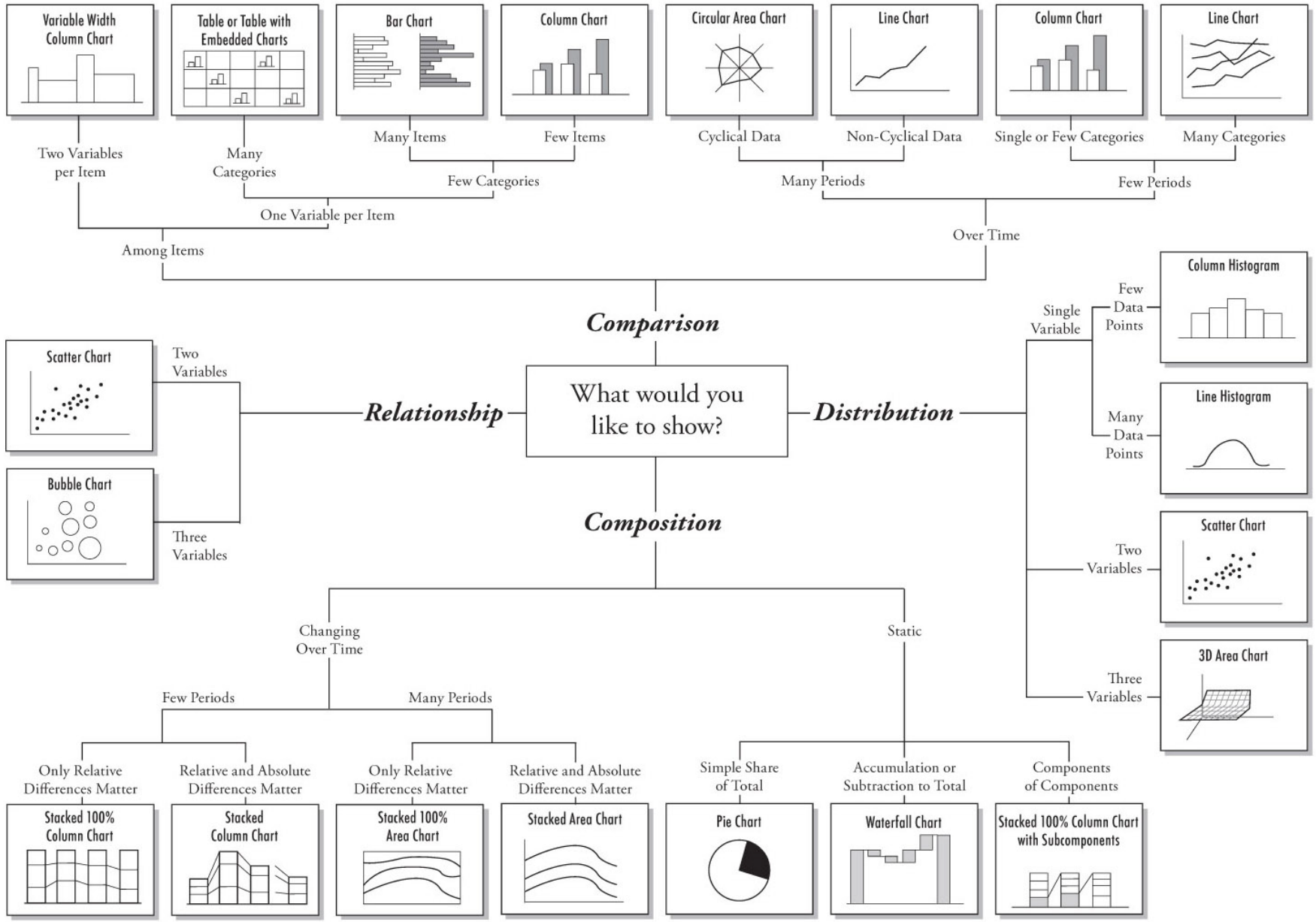
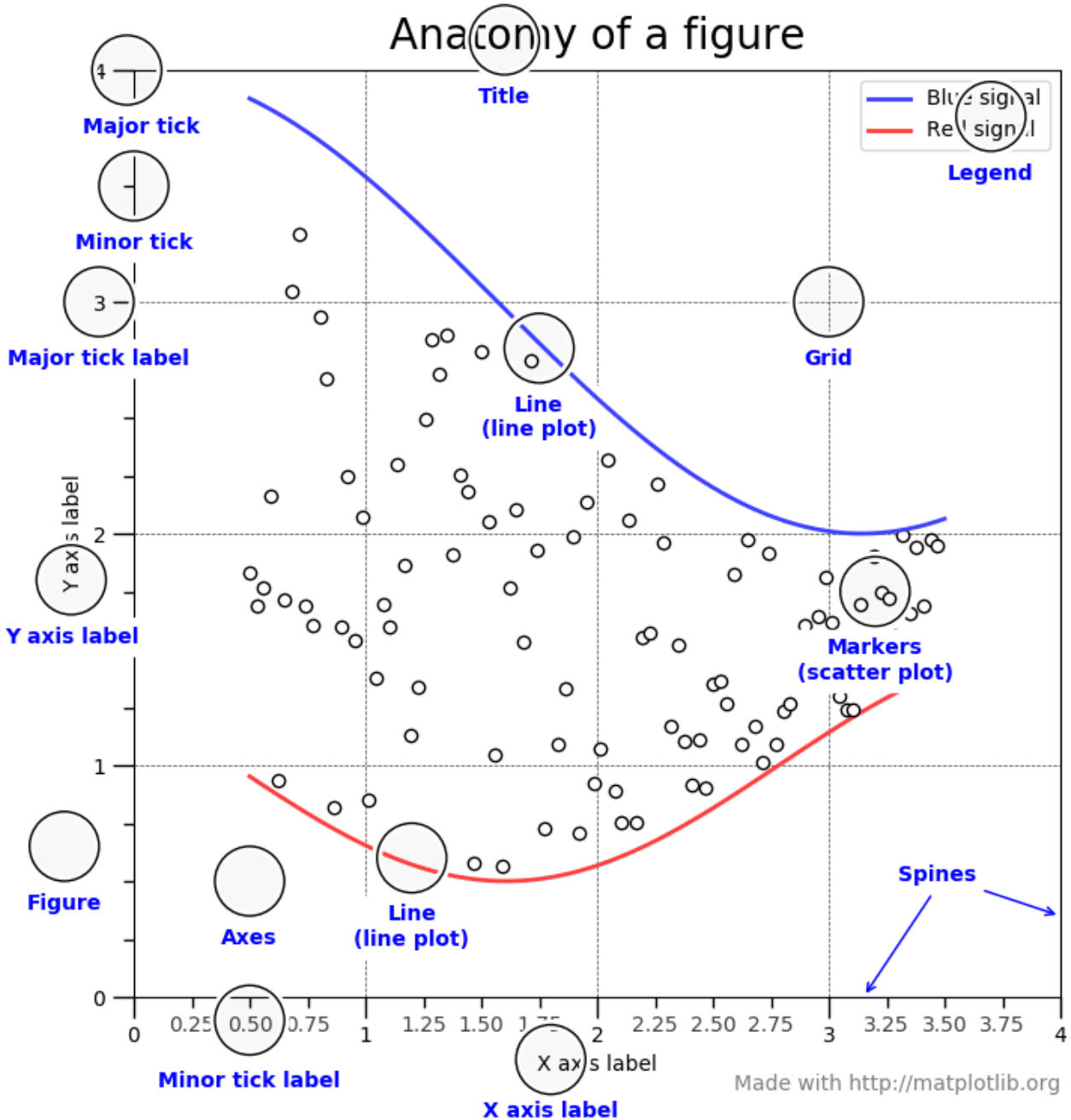


Chart Suggestions—A Thought-Starter




Anatomy of a figure
Major tick
Title
Minor tick
Legend
Major tick label
Grid
Line
(line plot)
Y axis label
Markers
(scatter plot)
Figure
Axes
Line
(line plot)
Spines
Minor tick label
X axis label
Made with http://matplotlib.org

Two Libraries

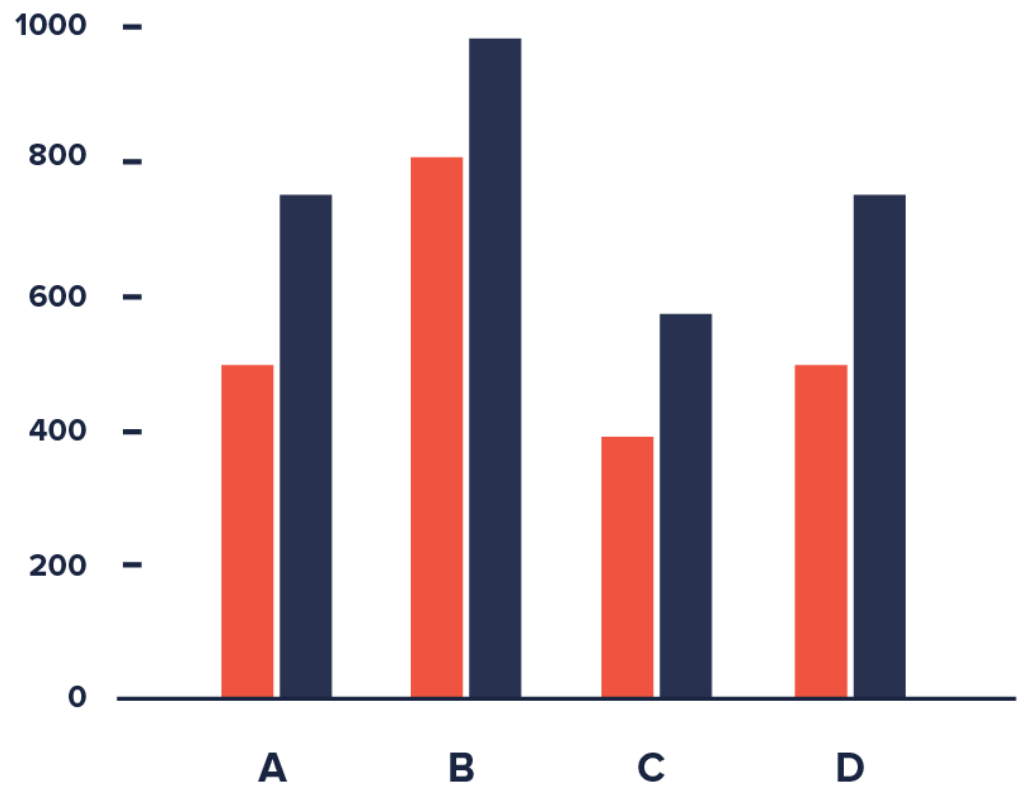
```
import matplotlib.pyplot as plt
import seaborn as sns
```

If you dont have them:

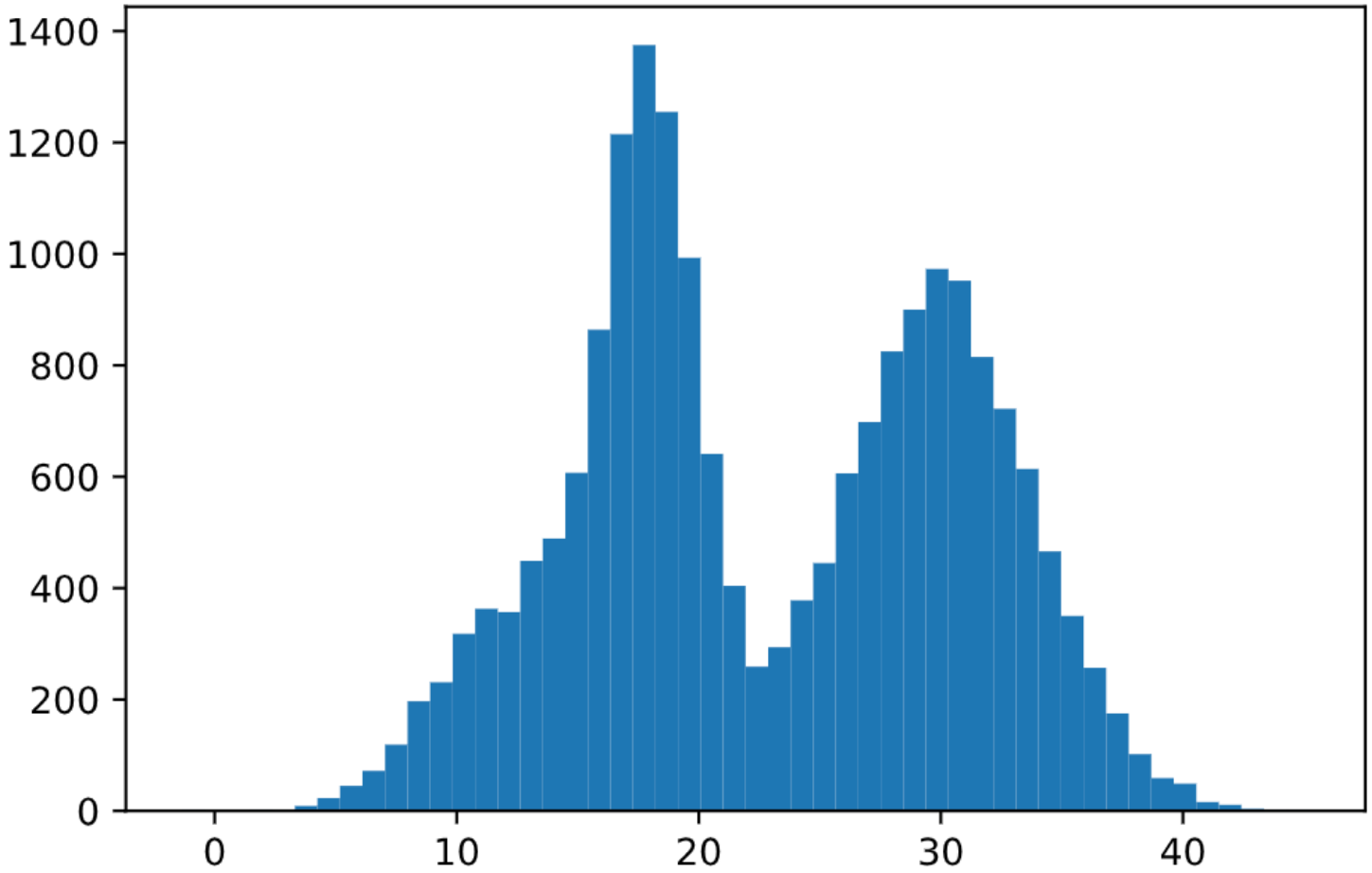
conda install matplotlib
conda install seaborn

1D DATA

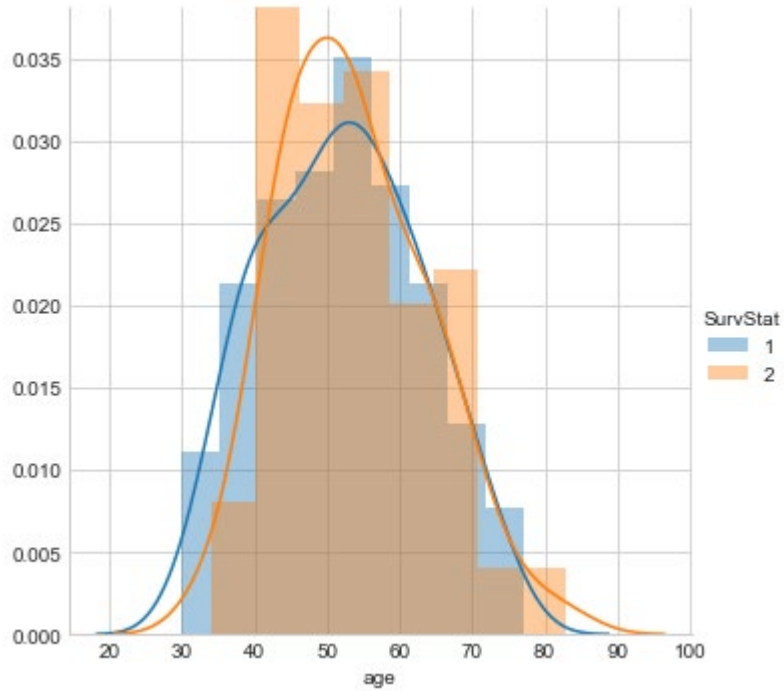
Bar Chart



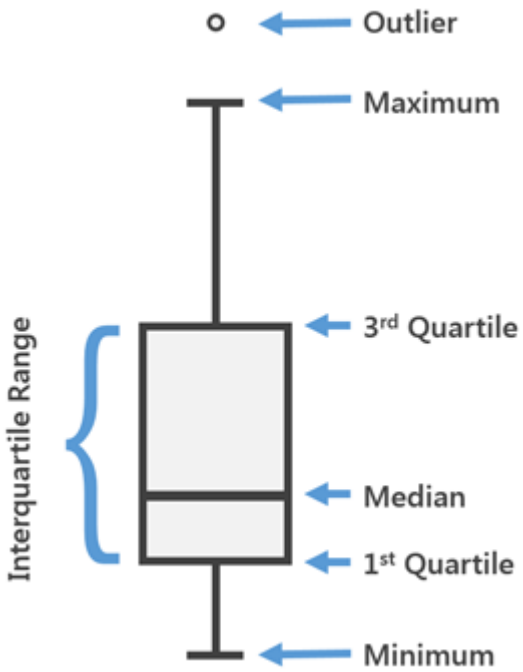
Histogram

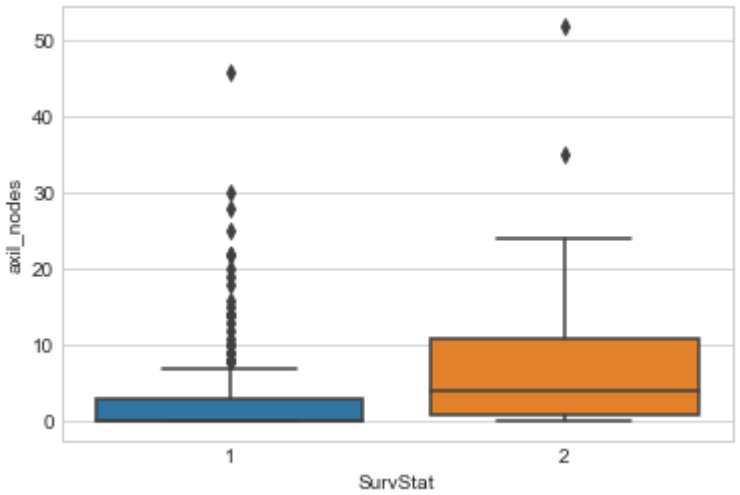


Distribution plot

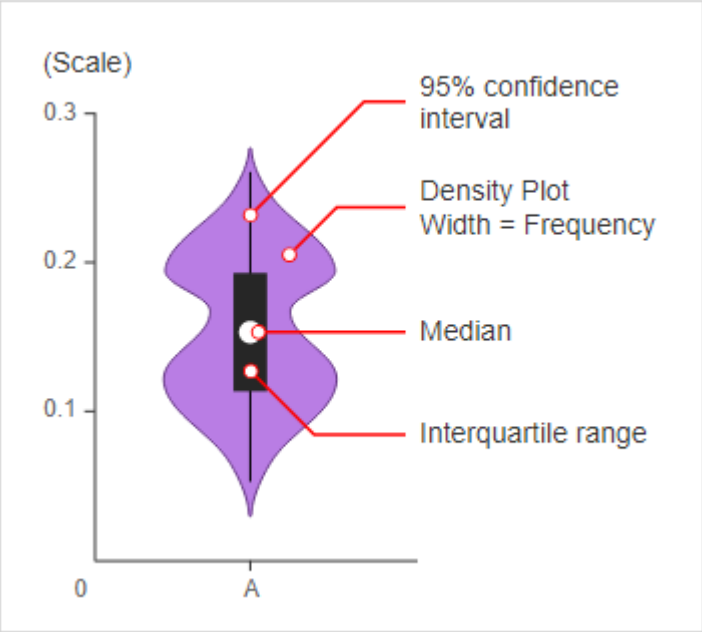


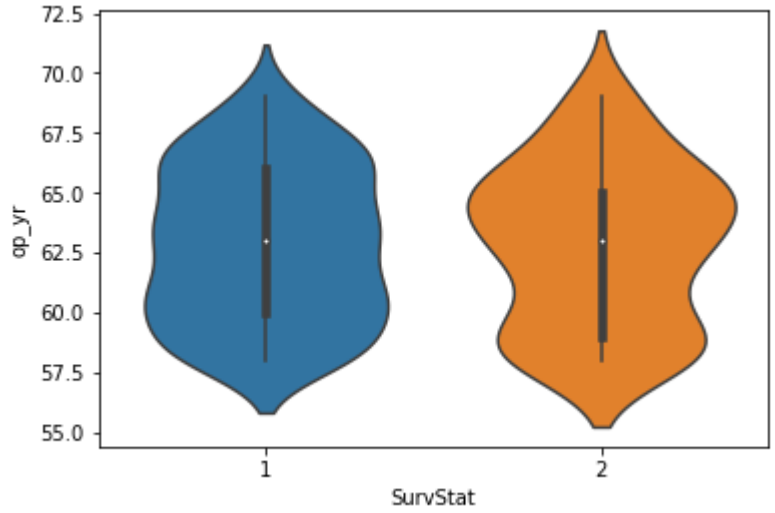
Box plot


Outlier
Maximum
3rd Quartile
Median
1st Quartile
Minimum
Interquartile Range


50
40
30
20
10
0
axil_nodes
1
2
SurvStat

Violin Plot

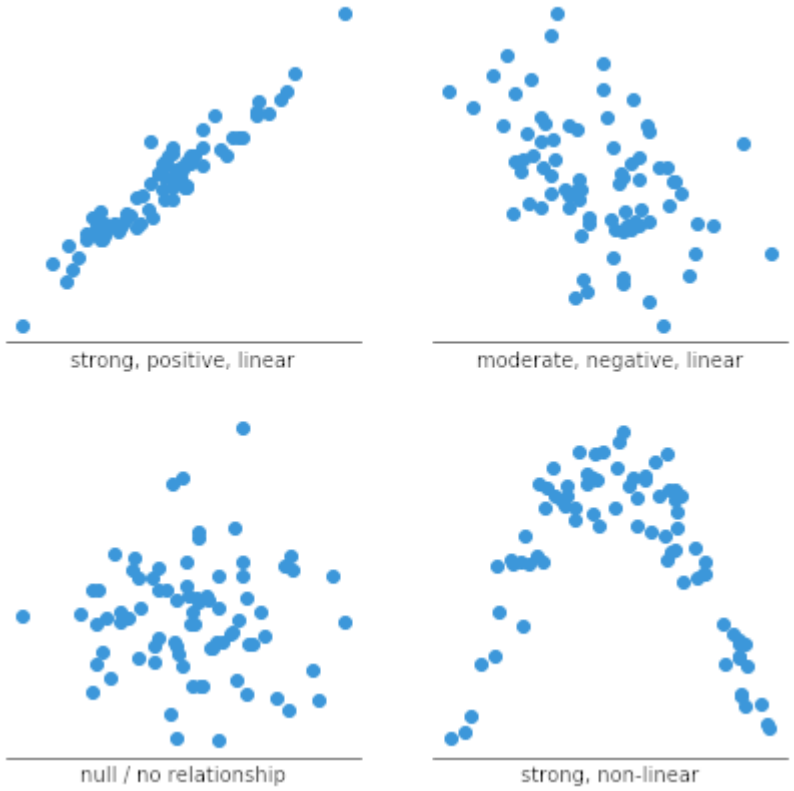

(Scale)
0.3
0.2
0.1
0
A
95% confidence interval
Density Plot
Width = Frequency
Median
Interquartile range

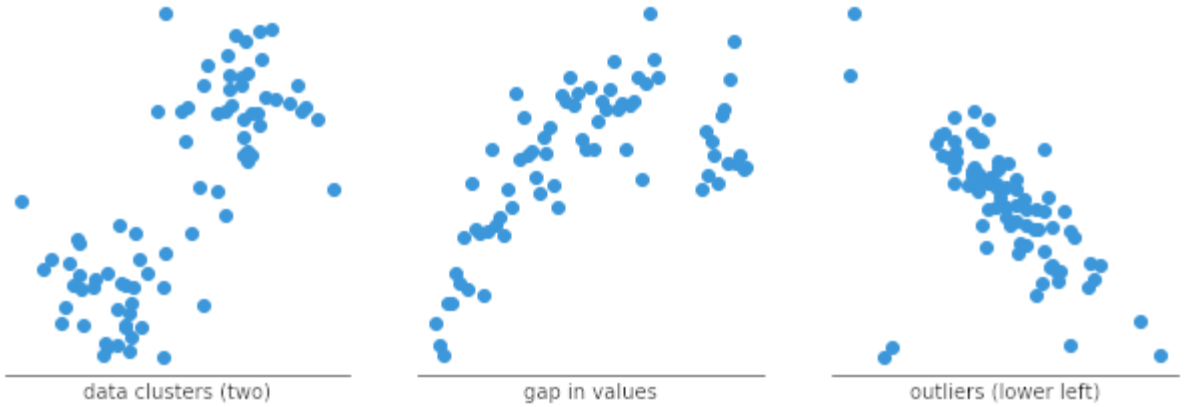

72.5
70.0
67.5
65.0
62.5
60.0
57.5
55.0
op_yr
1
2
SurvStat

2D DATA

Scatter plot

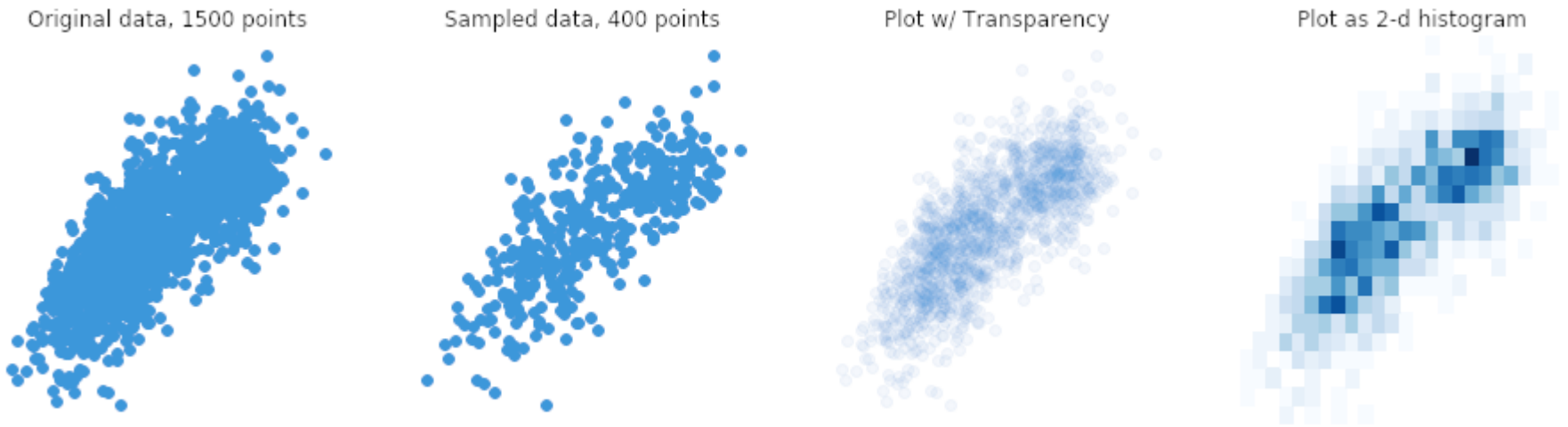
When you should use a scatter plot





Common issues

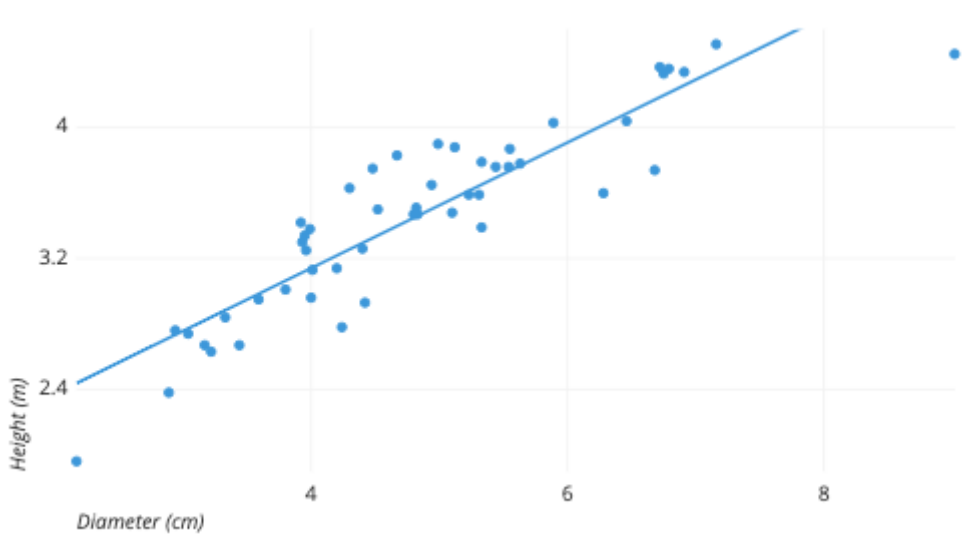
Overplotting



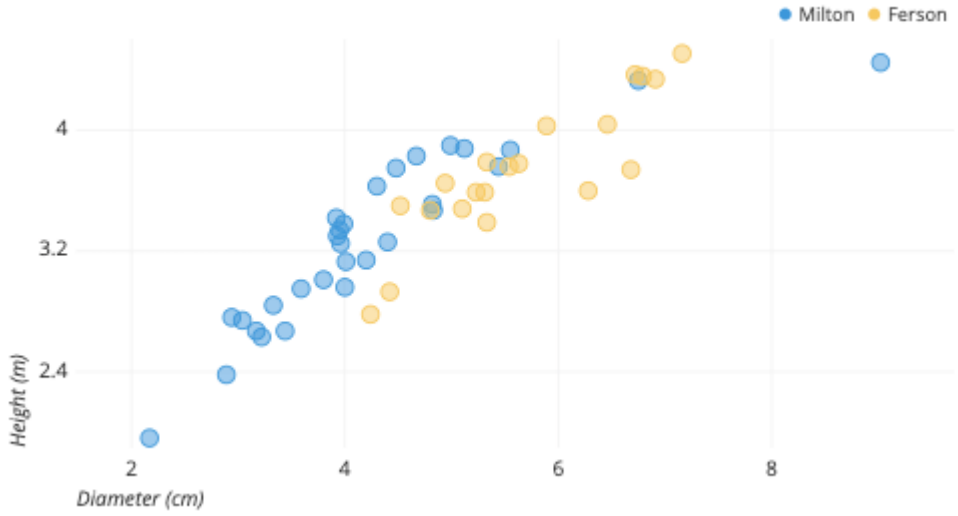
Interpreting correlation as causation

Common scatter plot options

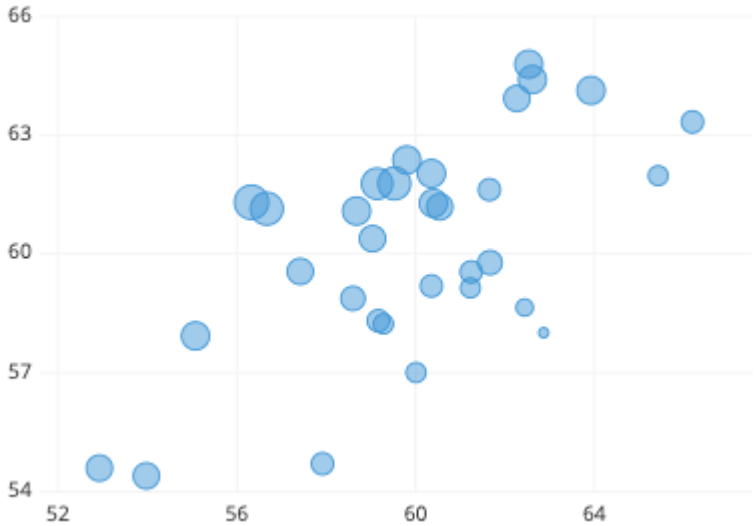
Add a trend line

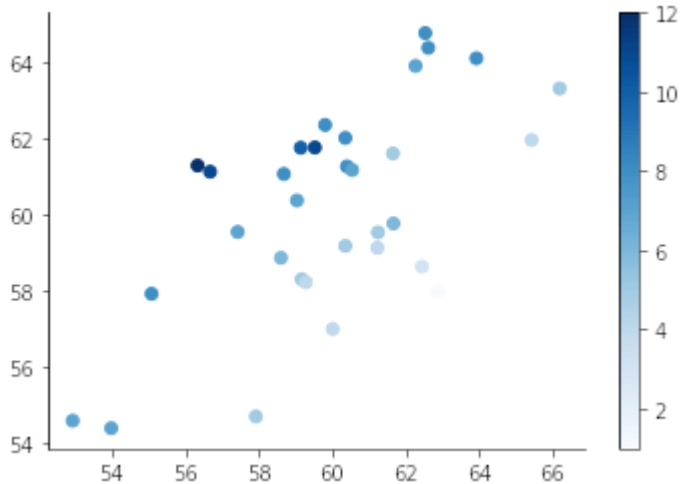


Categorical third variable

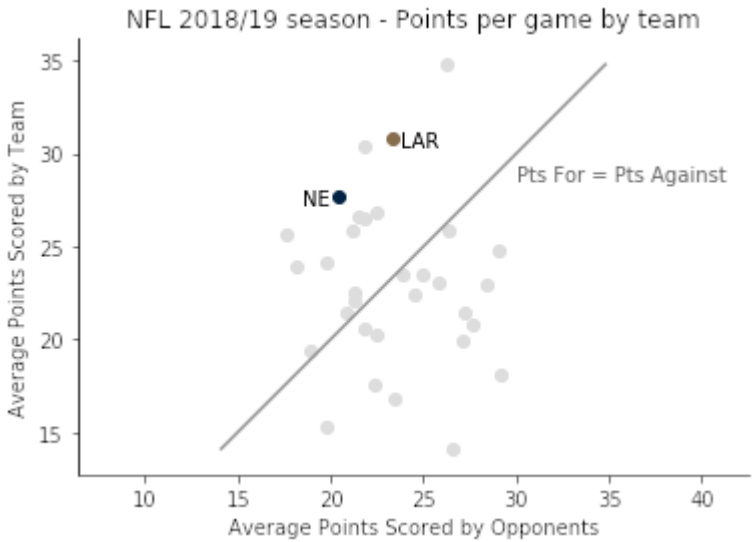


Numeric third variable

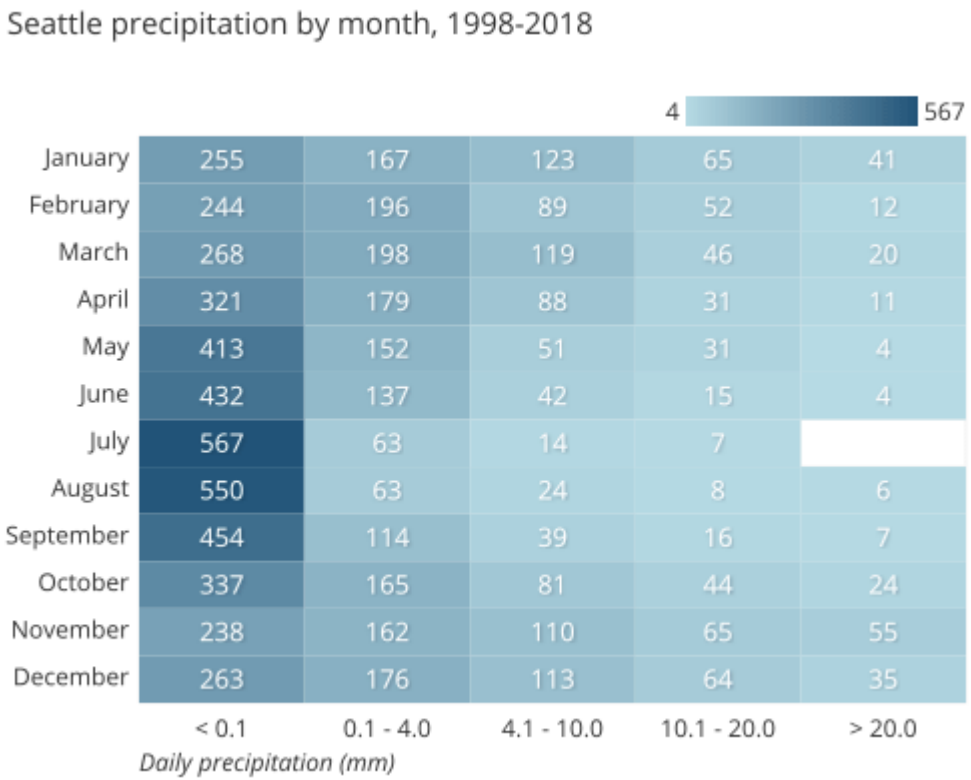


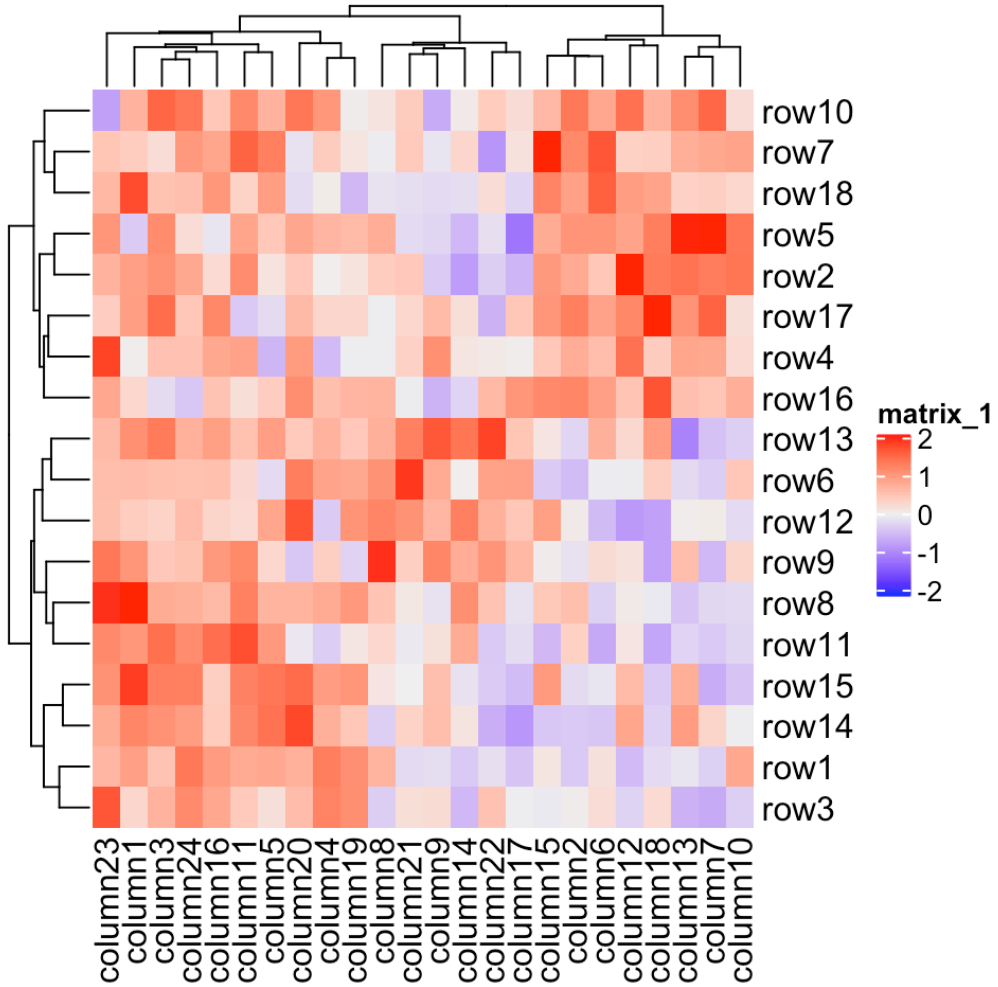


Highlight using annotations and color

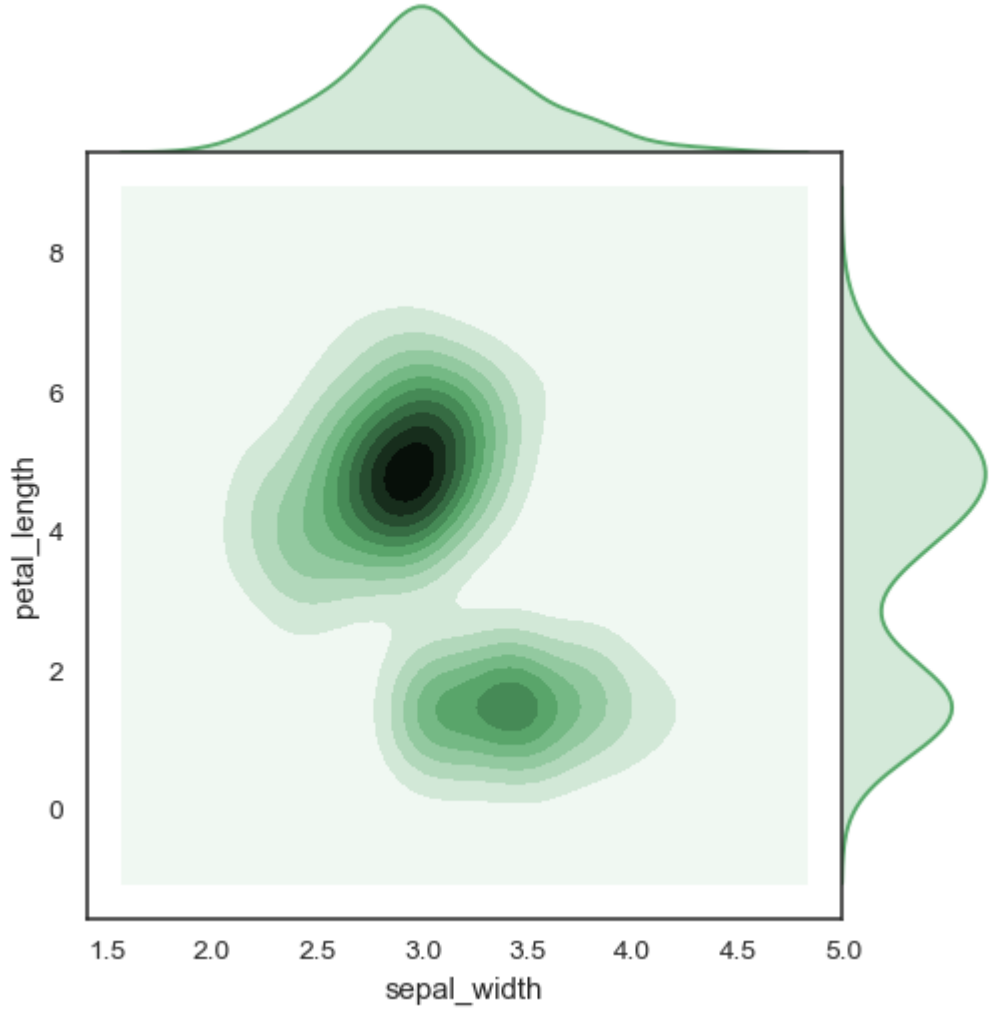


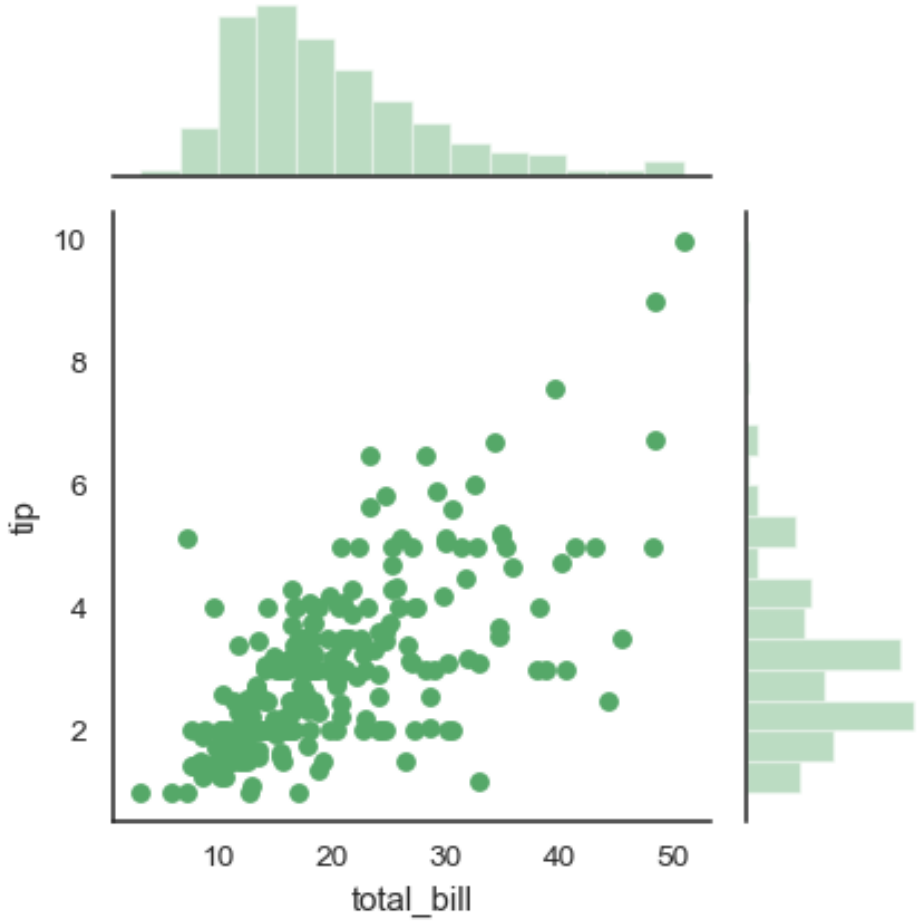
Heatmap (density, or 2D histogram)



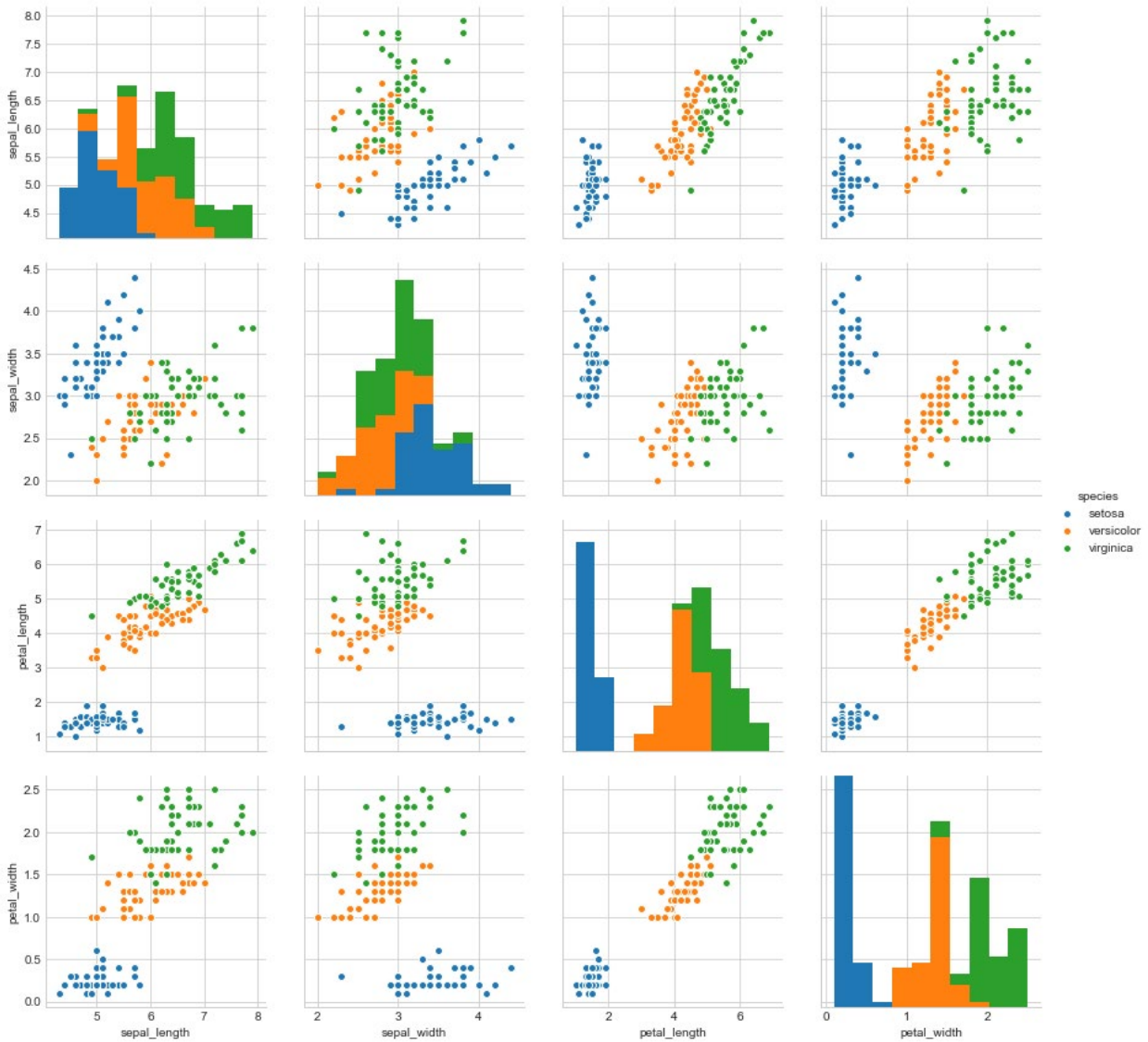


Joint plot


petal_length
sepal_width
0
2
4
6
8
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

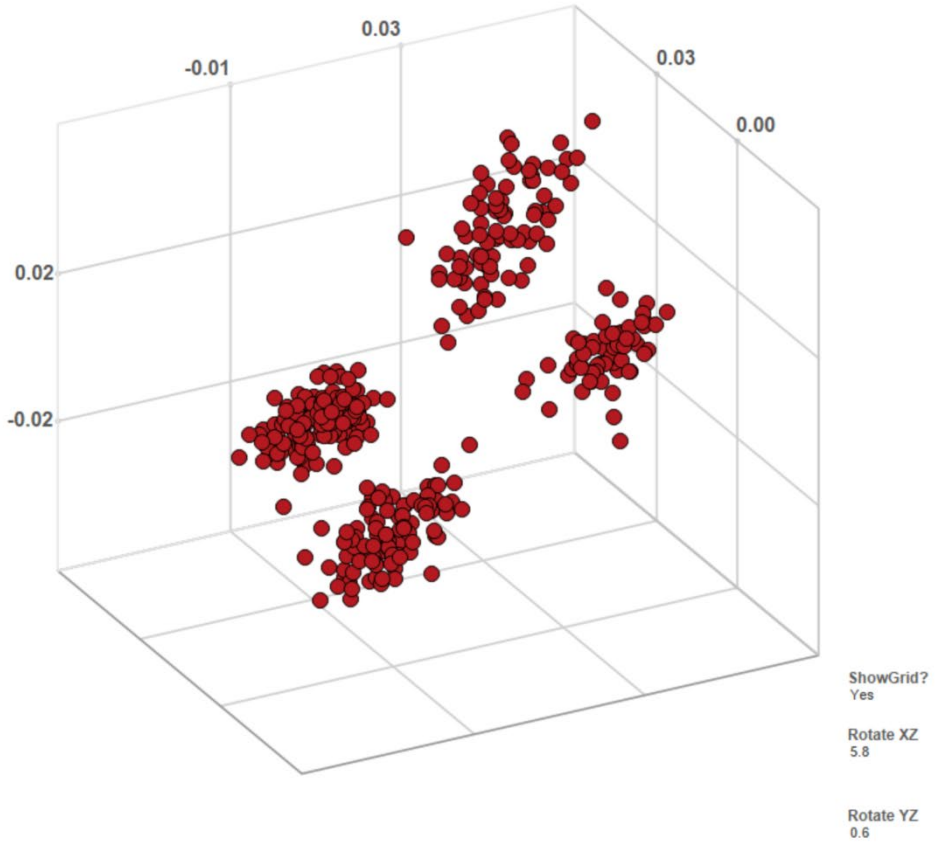

tip
total_bill
2
4
6
8
10
10
20
30
40
50

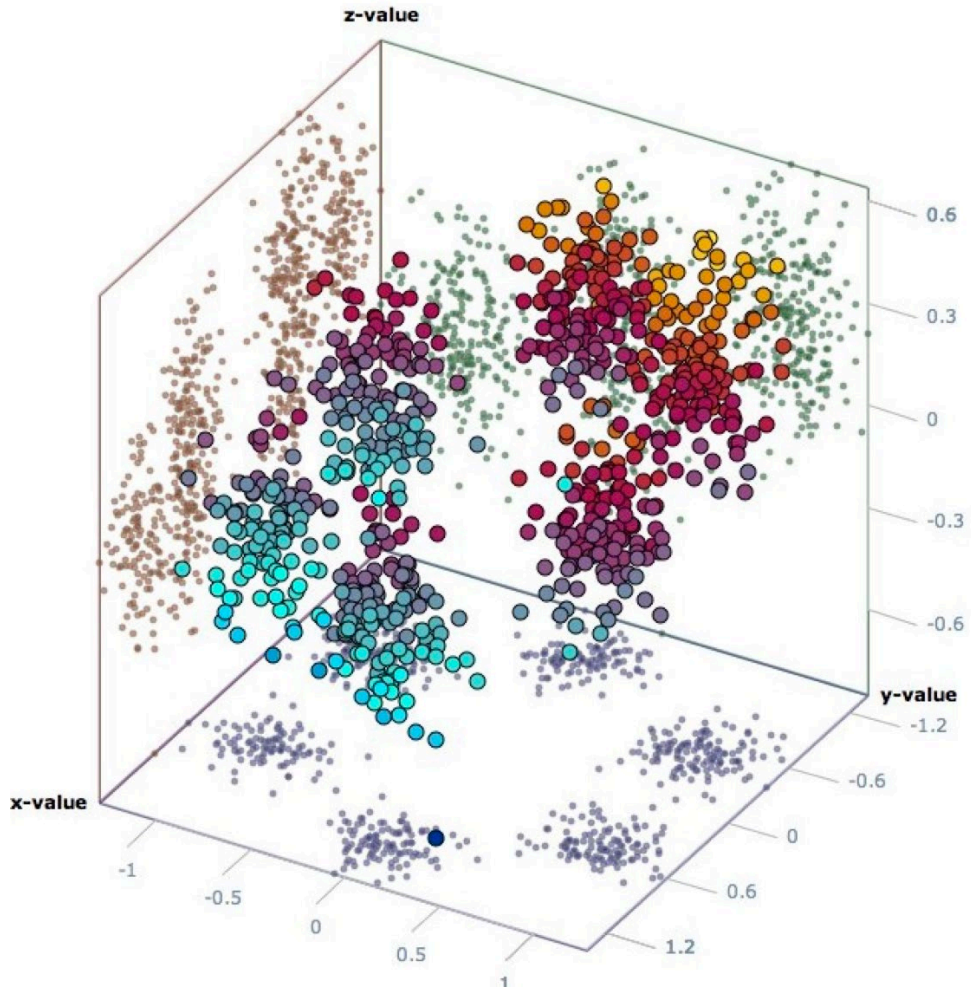
Pair plots


sepal_length
sepal_width
petal_length
petal_width
species
setosa
versicolor
virginica

3D DATA and beyond

3D Scatter


-0.01
0.03
0.03
0.00
0.02
-0.02
ShowGrid?
Yes
Rotate XZ
5.8
Rotate YZ
0.6


z-value
0.6
0.3
0
-0.3
-0.6
y-value
-1.2
-0.6
0
0.6
1.2
x-value
-1
-0.5
0
0.5
1

High-Throughput Experimentation

Assay Plates

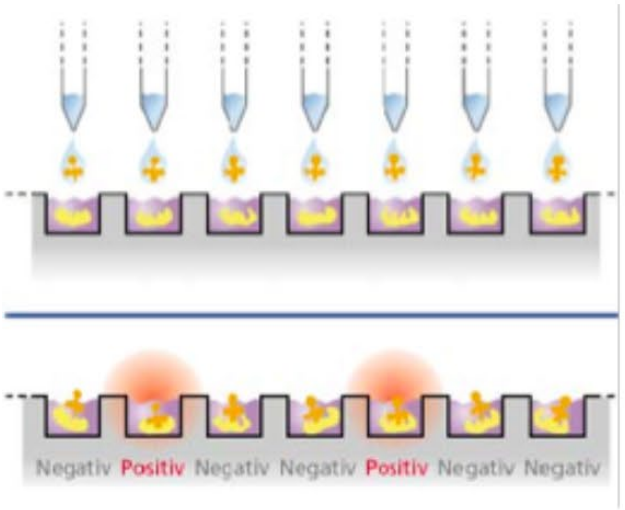
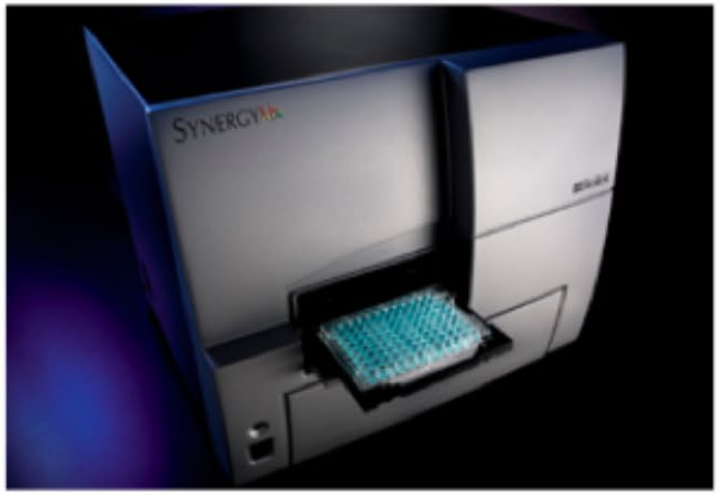
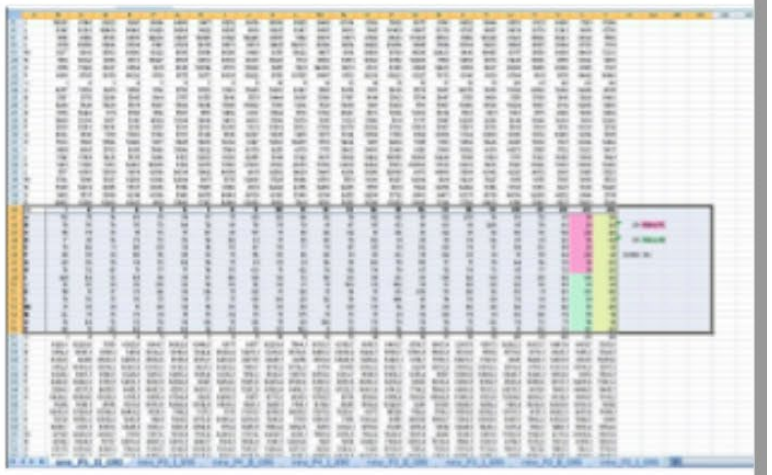


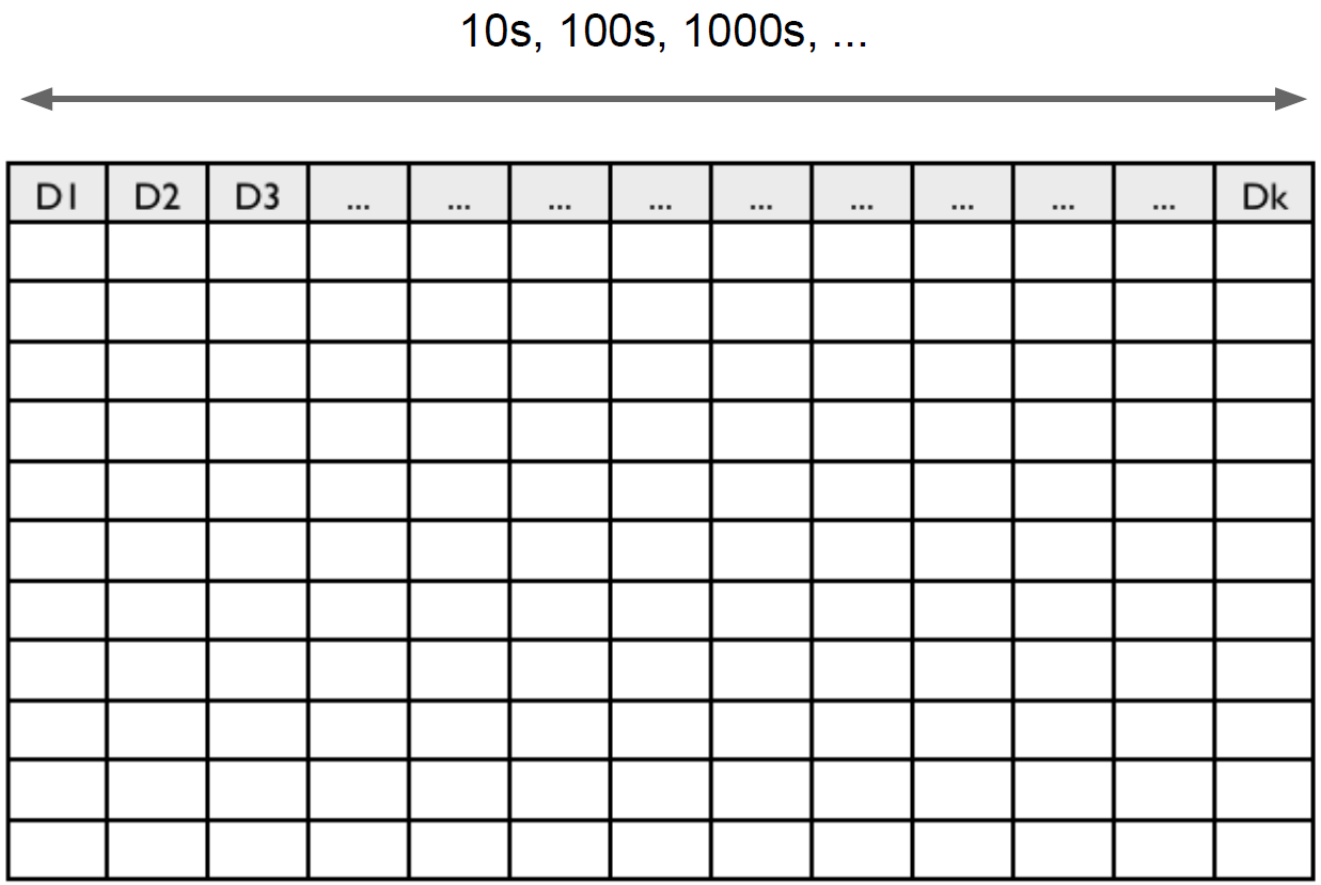
Plate Reader



Data



High-Dimensional Data Visualization

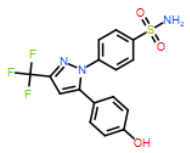
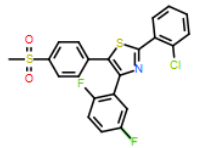
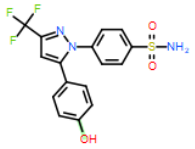
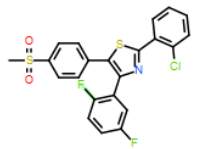
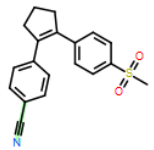
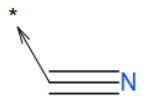
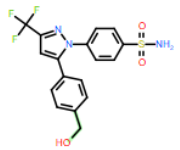
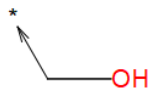
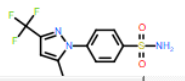


- Set of visual features very limited
- Resolution very limited
- Ability to make sense of it very limited!

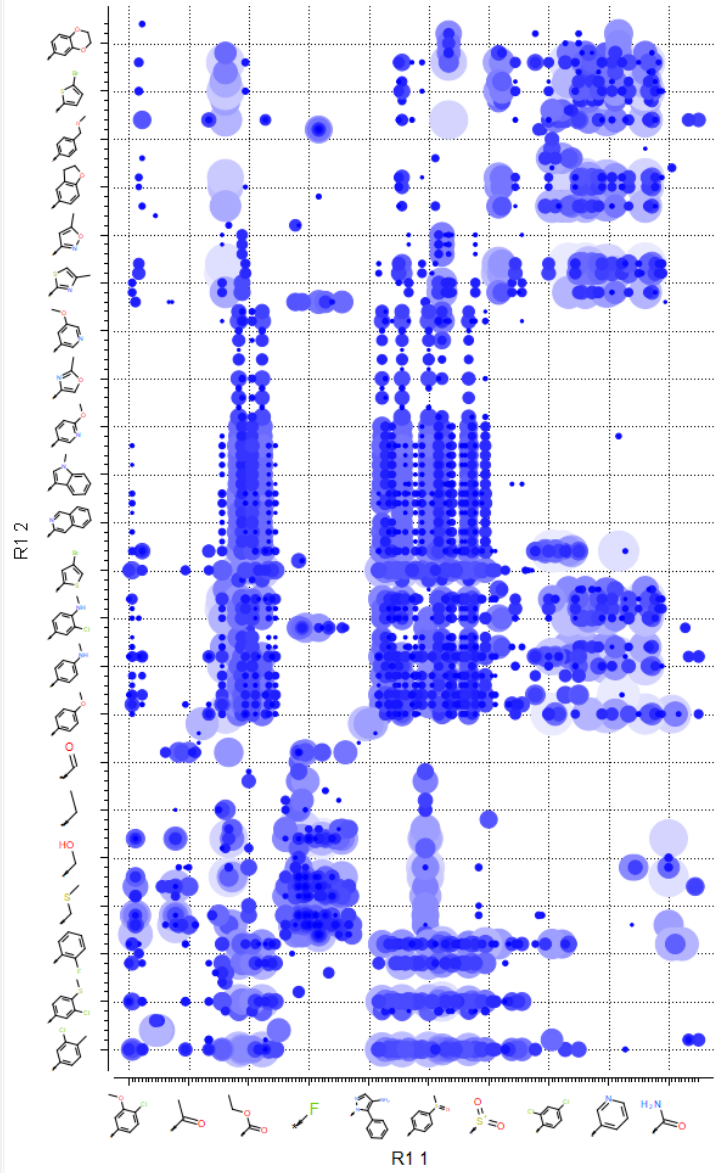
webinar_sar.icb Molsoft icm 3.8-7 [Z:\trainingcourse\webinar_sar\webinar_sar.icb] (5 tables)

File Tools Edit View Bioinfo Homology Chemistry Docking MolMechanics Windows Help

COX2 | COX2_scaffold | T_sar_R1_R4 | COX2_SALI | COX2_pairs

	nol1	R1 1	ix2
1		smiles 	247
2		smiles 	36
3		smiles 	253
4		smiles 	37
5		smiles 	156
6		smiles 	253
		smiles 	247

Header | Filter | R1_1-R1_2



ACTIVITY_1..

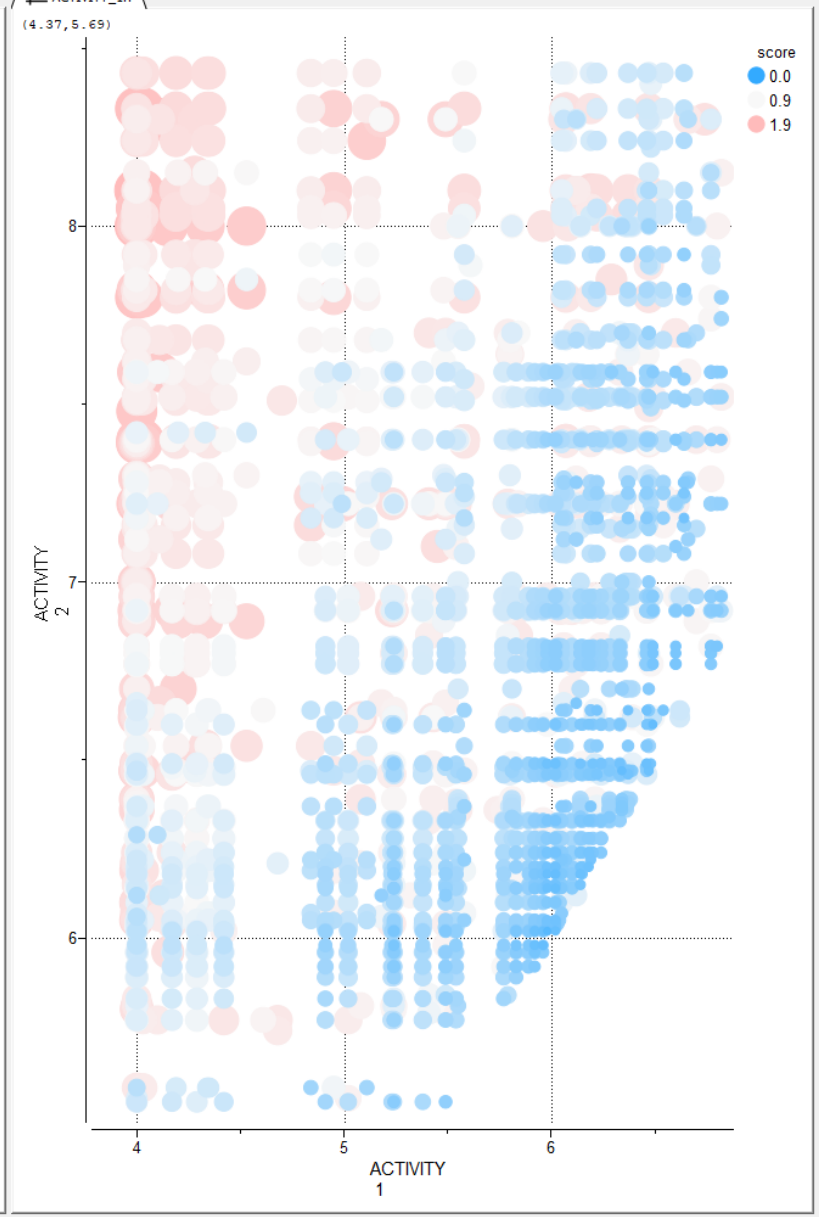
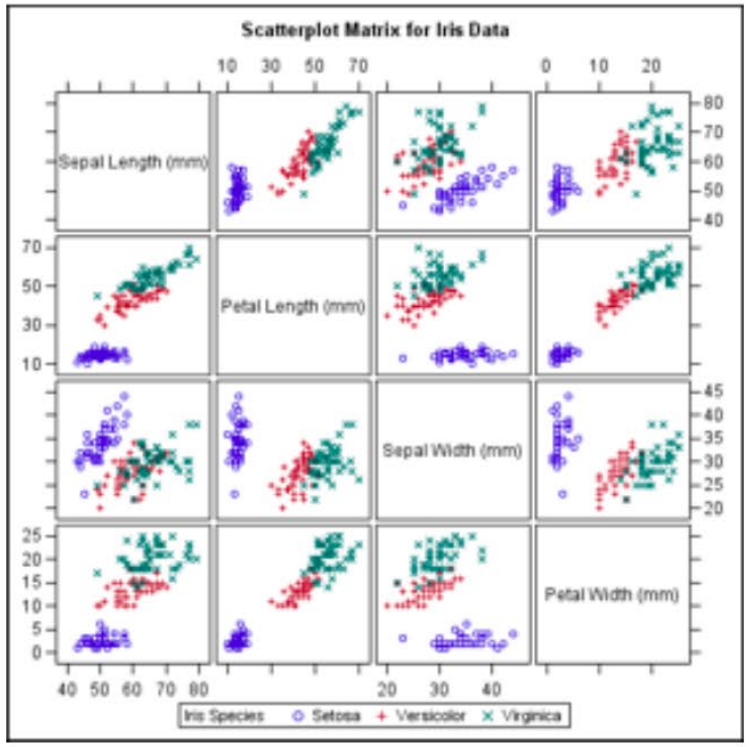


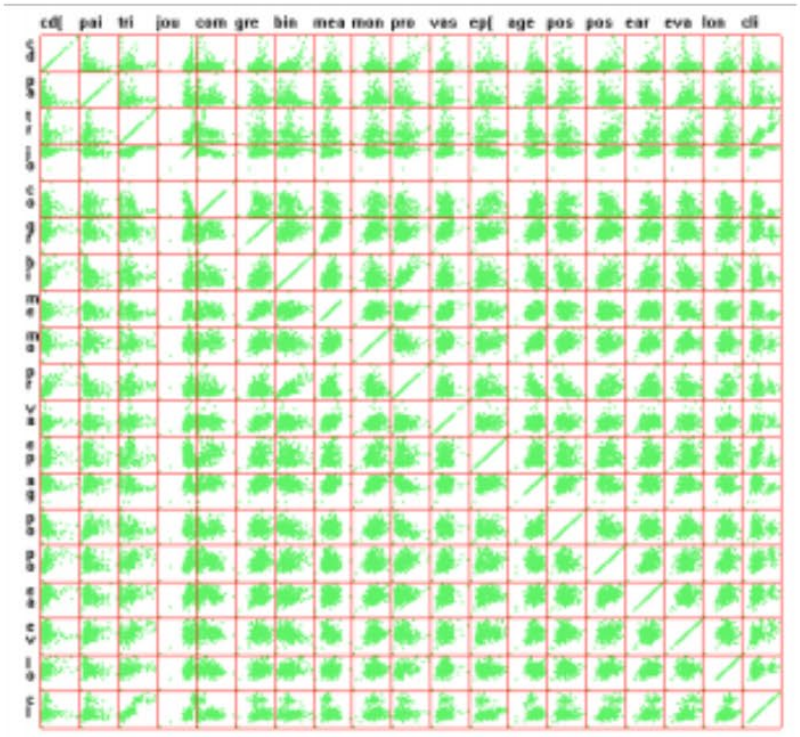
table: 3206 rows, 11 columns (2 selected columns, 3206 non-empty rows)

Example: Scatter Plot Matrix

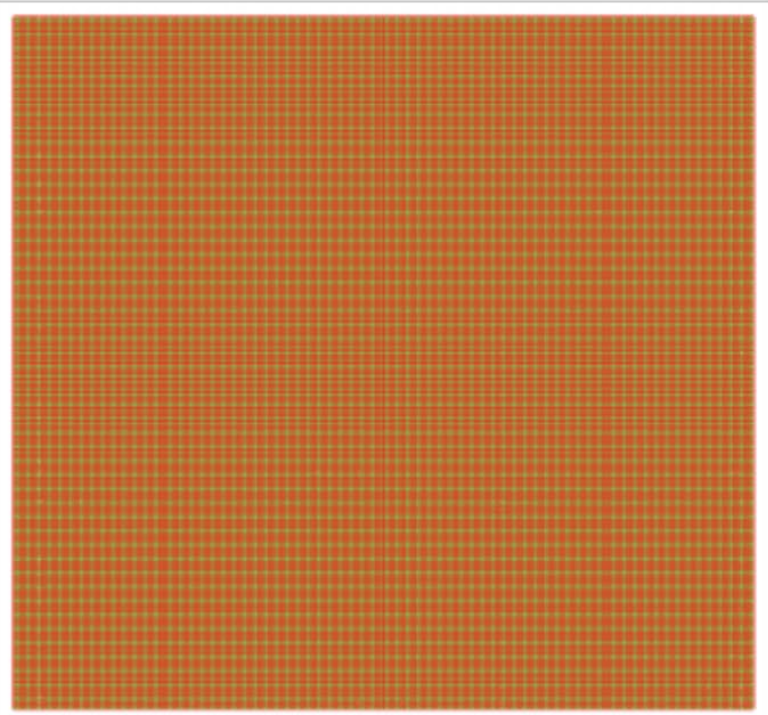
4 dimensions



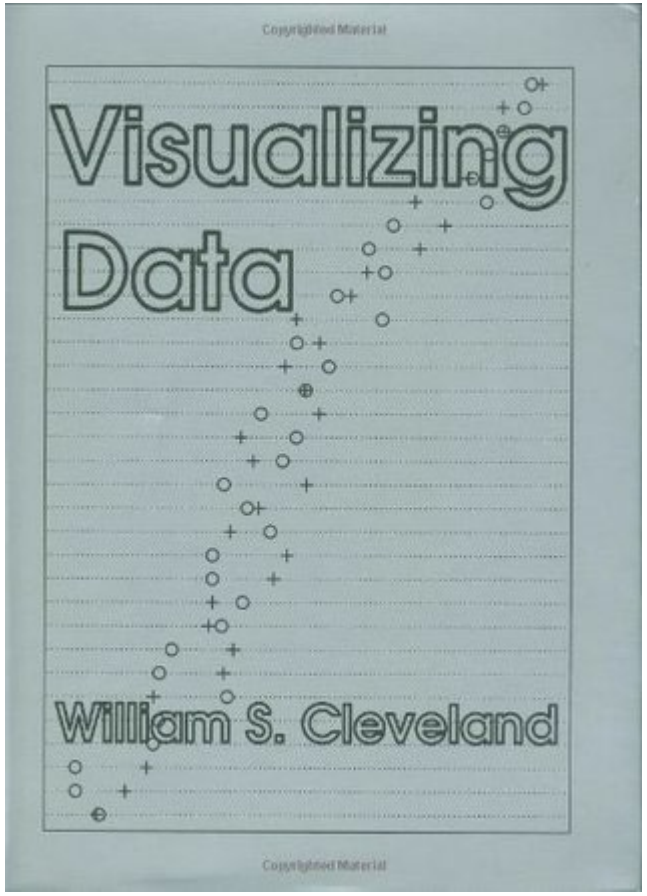
20 dimensions

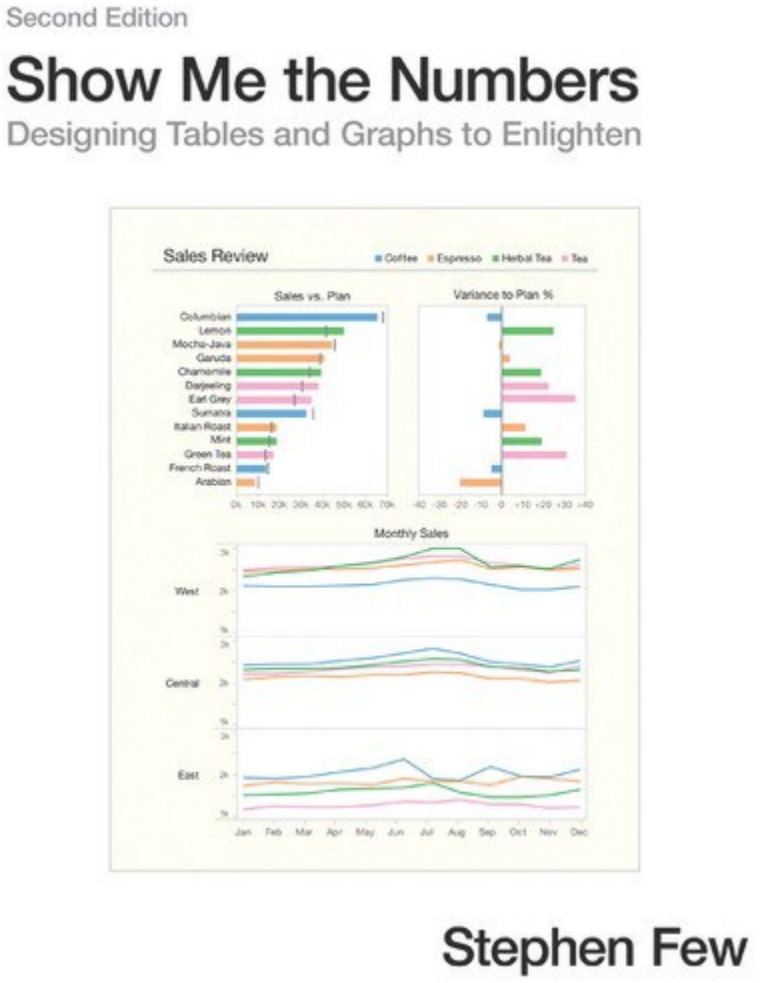


100 dimensions



If you want to learn more…





Show Me the Numbers
Designing Tables and Graphs to Enlighten, 2nd Ed.