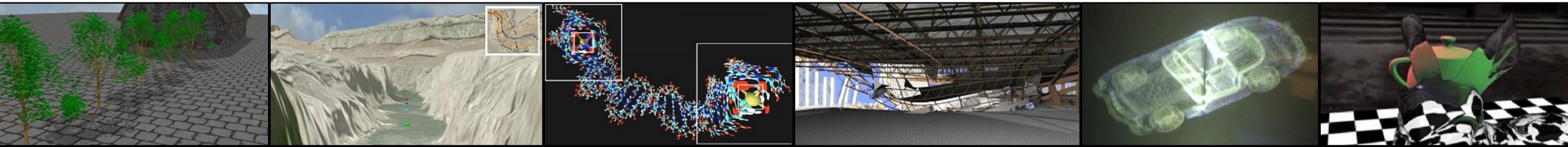
CIS 4930/6930-002
DATA VISUALIZATION

Descriptive Statistics and Visualization

Ghulam Jilani Quadri
University of South Florida

Slide credits D.A. Forsyth



Reminders

4/9/2018 – Project 7 Due

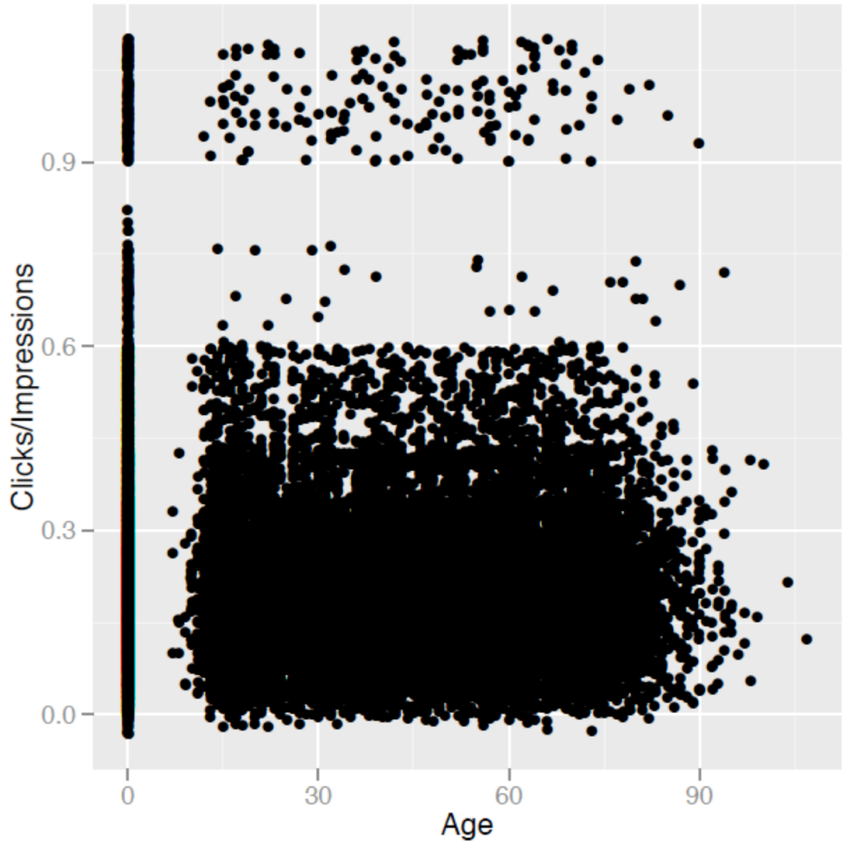
4/11/2018 – Paper 4 Review Due

4/16/2018 – Project 7 Peer Review Due

4/23/2018 – Project 8 Due

Problem #1:

We have too many data points to show

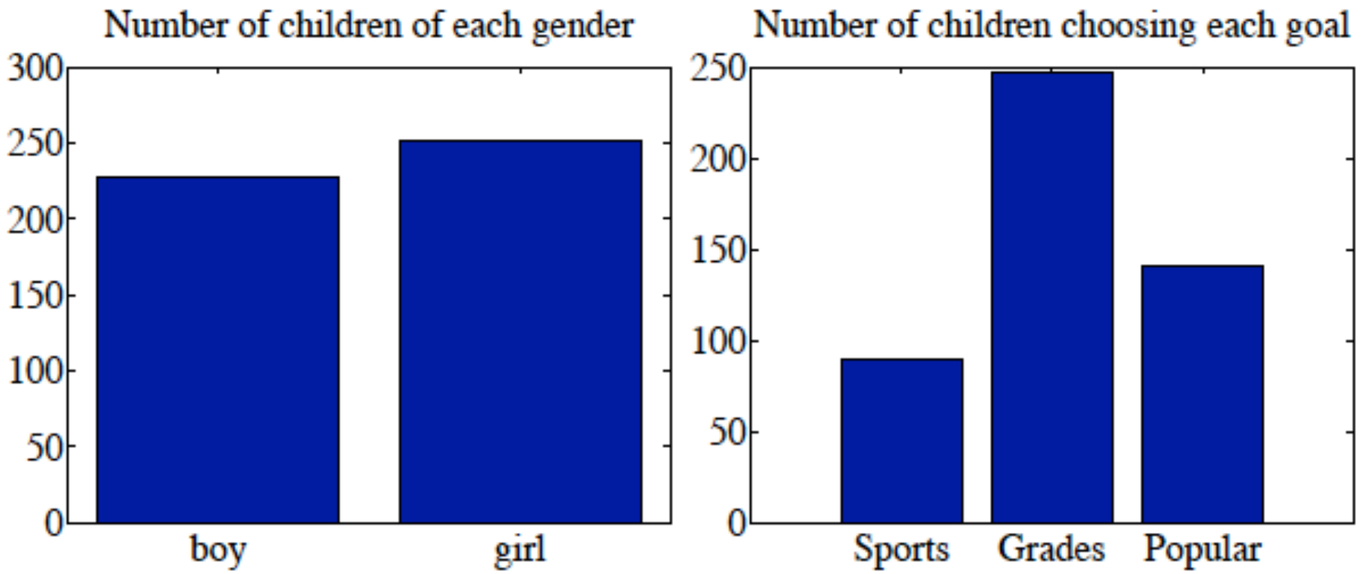


HISTOGRAMS

Bar chart-based visualization that allows evaluating distribution of values.

CATEGORICAL DATA

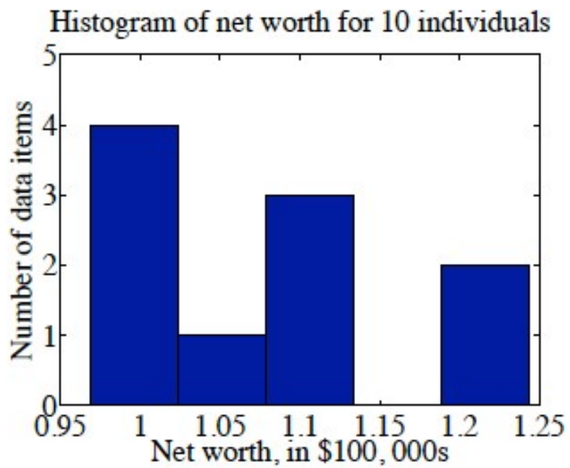
Gender	Goal	Gender	Goal
boy	Sports	girl	Sports
boy	Popular	girl	Grades
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	girl	Grades
girl	Popular	girl	Sports
girl	Grades	girl	Popular
girl	Sports	girl	Grades
girl	Sports	girl	Sports

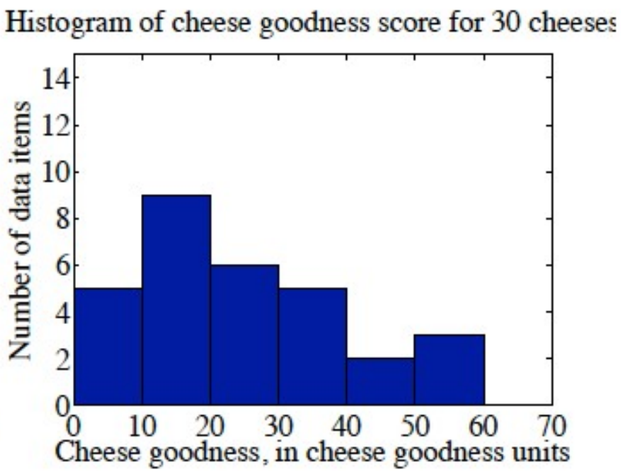


Continuous Data Histograms

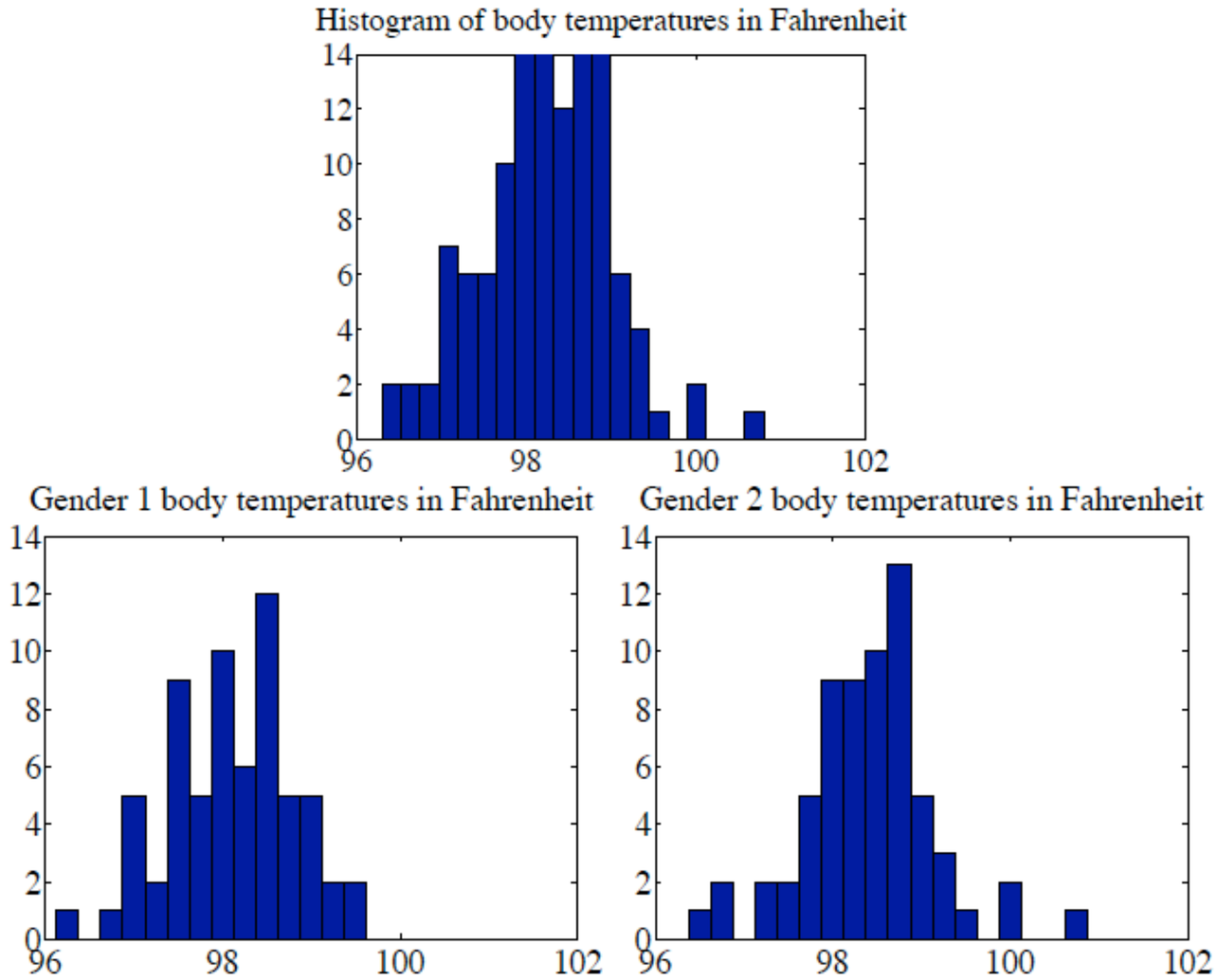
Index	net worth
1	100, 360
2	109, 770
3	96, 860
4	97, 860
5	108, 930
6	124, 330
7	101, 300
8	112, 710
9	106, 740
10	120, 170

Index	Taste score	Index	Taste score
1	12.3	11	34.9
2	20.9	12	57.2
3	39	13	0.7
4	47.9	14	25.9
5	5.6	15	54.9
6	25.9	16	40.9
7	37.3	17	15.9
8	21.9	18	6.4
9	18.1	19	18
10	21	20	38.9





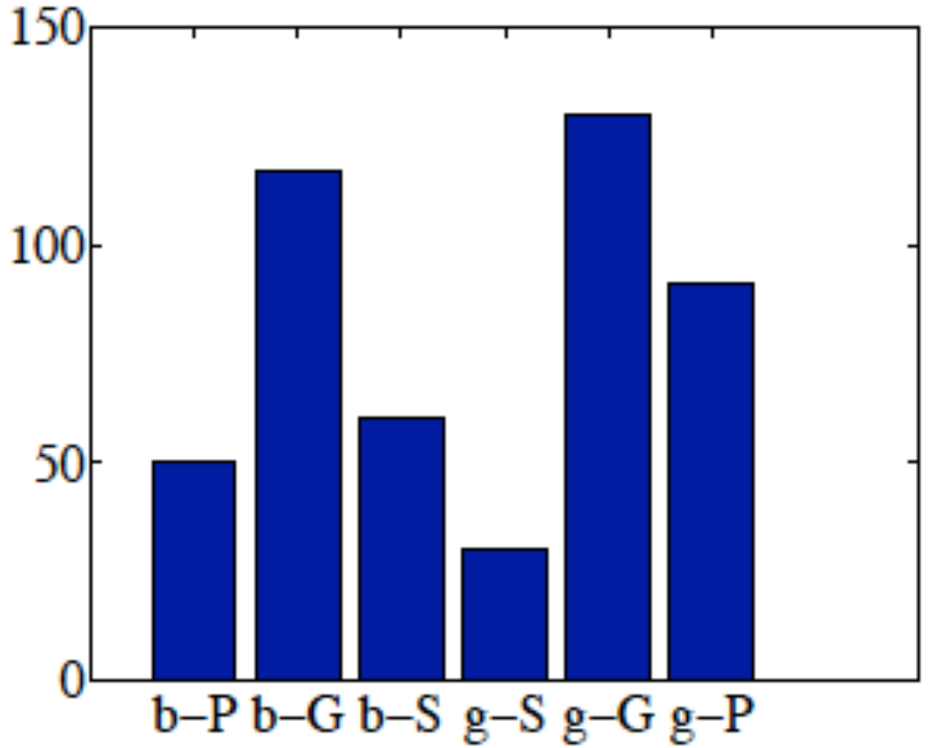
Conditional Histograms



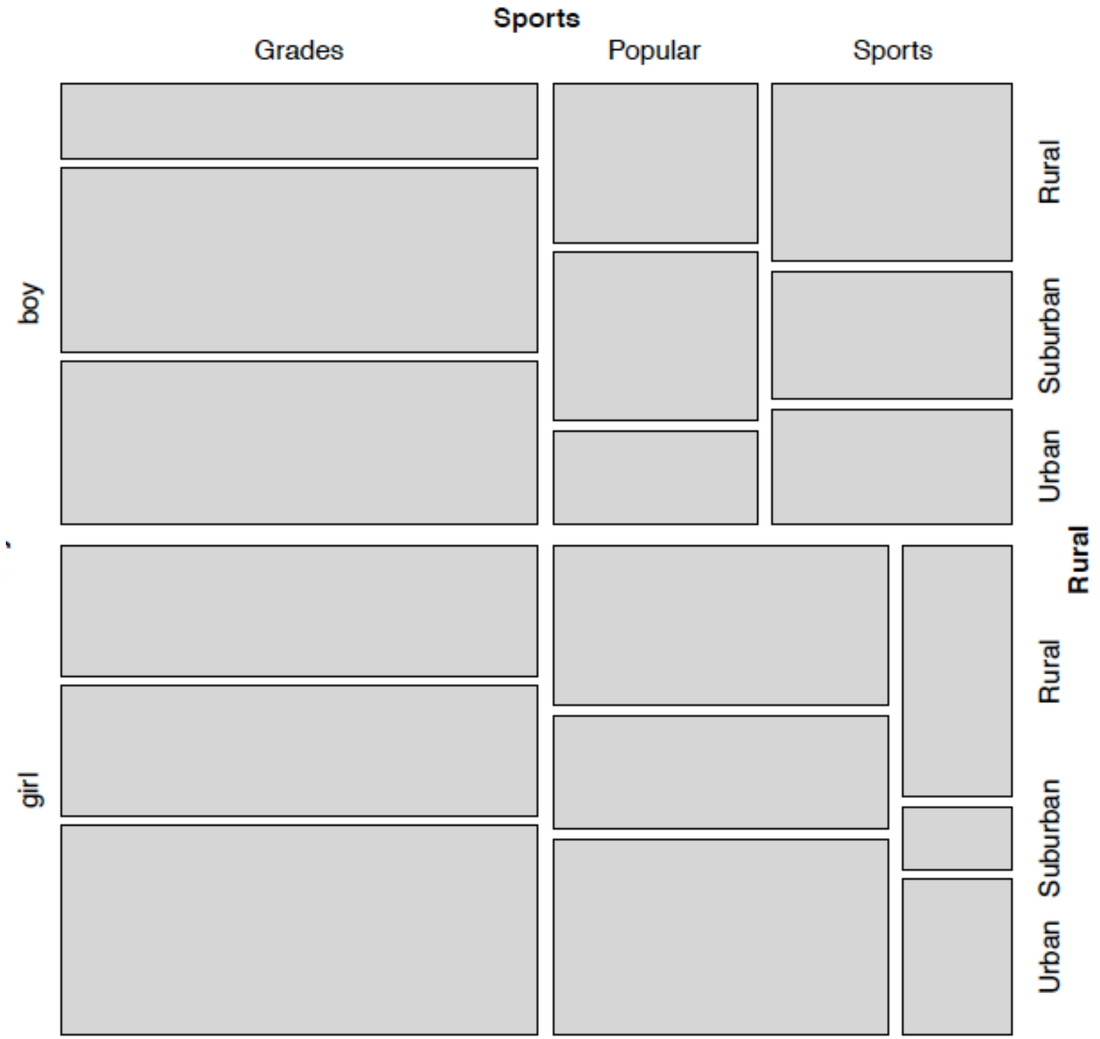
2D Data

CATEGORICAL DATA

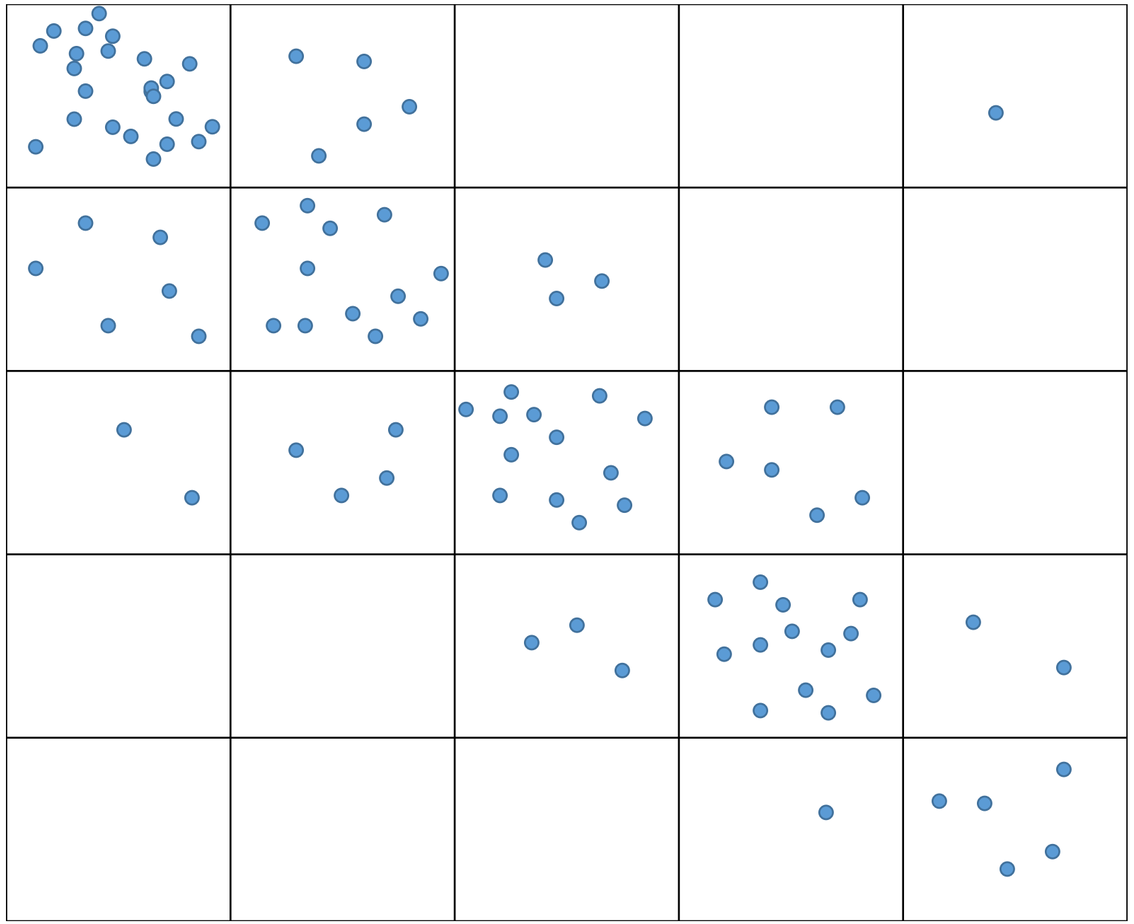
Gender	Goal	Gender	Goal
boy	Sports	girl	Sports
boy	Popular	girl	Grades
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	girl	Grades
girl	Popular	girl	Sports
girl	Grades	girl	Popular
girl	Sports	girl	Grades
girl	Sports	girl	Sports



Mosaic Plots



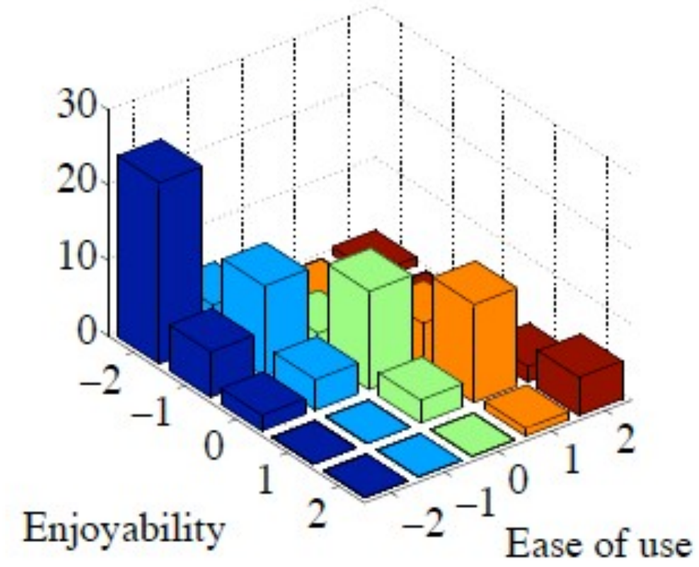
ORDINAL DATA

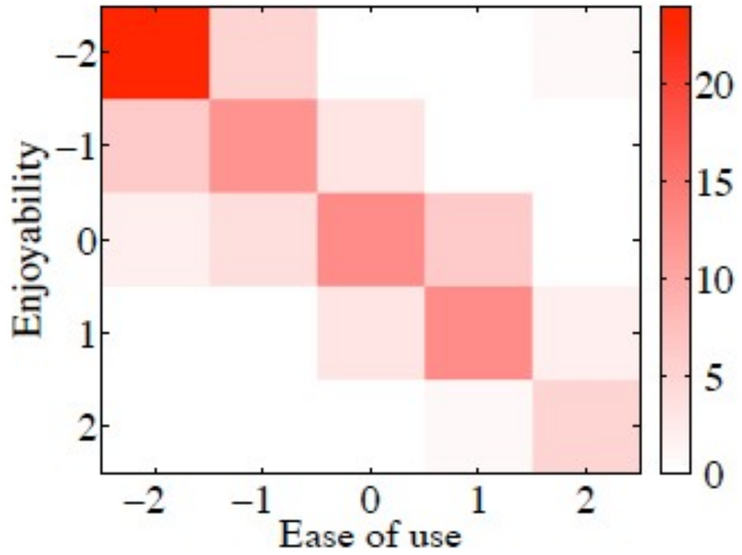


	-2	-1	0	1	2
-2	24	5	0	0	1
-1	6	12	3	0	0
0	2	4	13	6	0
1	0	0	3	13	2
2	0	0	0	1	5

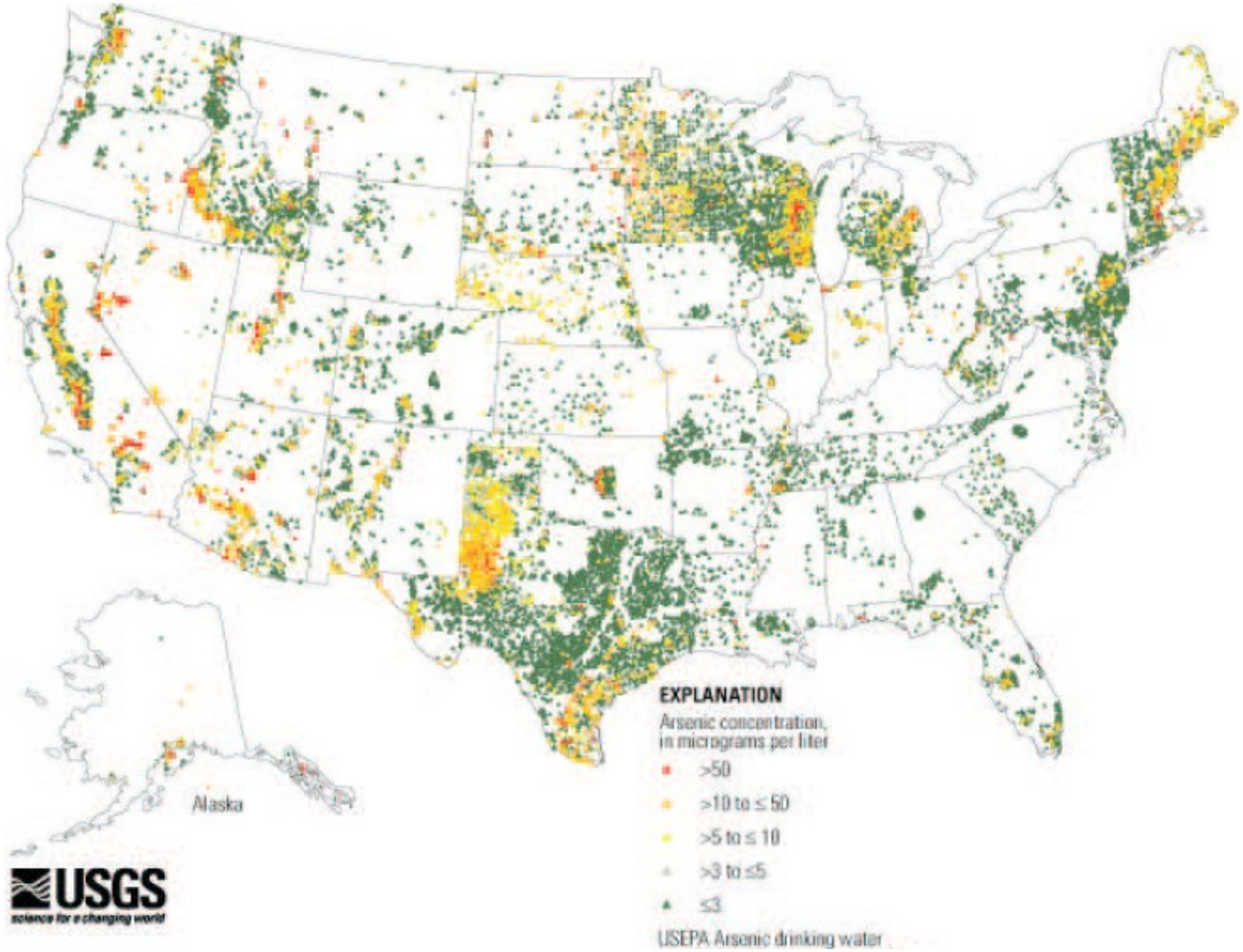
ORDINAL DATA

Counts of user responses for a user interface

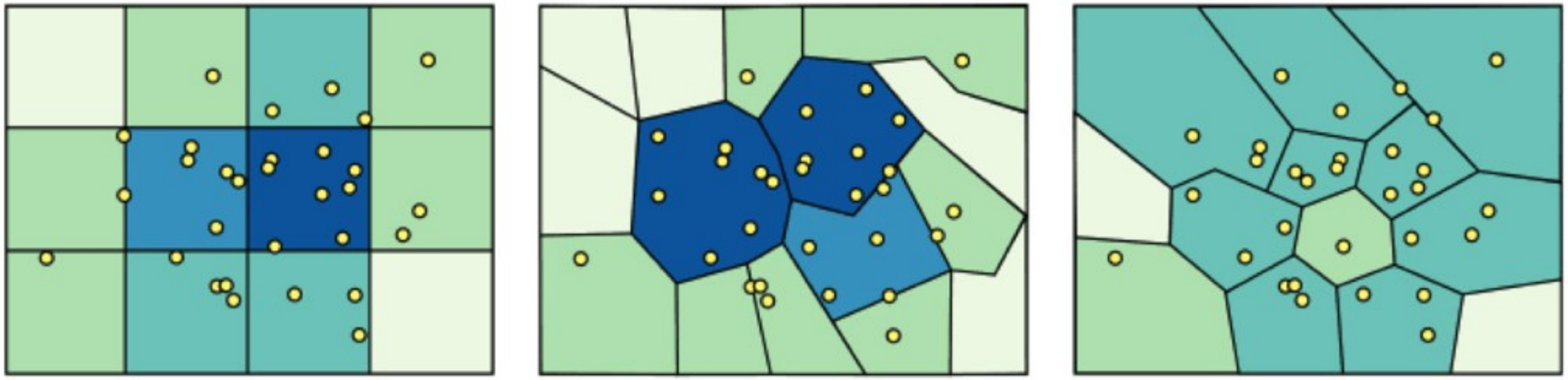




ARSENIC IN WELL WATER



SPATIAL AGGREGATION



MODIFIABLE AREAL UNIT PROBLEM

in cartography, changing the boundaries of the regions used to analyze data can yield dramatically different results

Modeling Data

SUMMARY STATISTICS – MEAN

Definition: 3.1 *Mean*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. Their mean is

$$\text{mean}(\{x\}) = \frac{1}{N} \sum_{i=1}^{i=N} x_i.$$

The average

The best estimate of the value of a new data point in the absence of any other information about it

SUMMARY STATISTICS - STANDARD DEVIATION

Definition: 3.2 *Standard deviation*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. The standard deviation of this dataset is is:

$$\text{std}(x_i) = \sqrt{\frac{1}{N}\sum_{i-1}^{i=N}(x_i - \text{mean}(\{x\}))^2} = \sqrt{\text{mean}(\{(x_i - \text{mean}(\{x\}))^2\})}.$$

Think of this as a scale

Average distance from mean

Standard Score (aka z score)

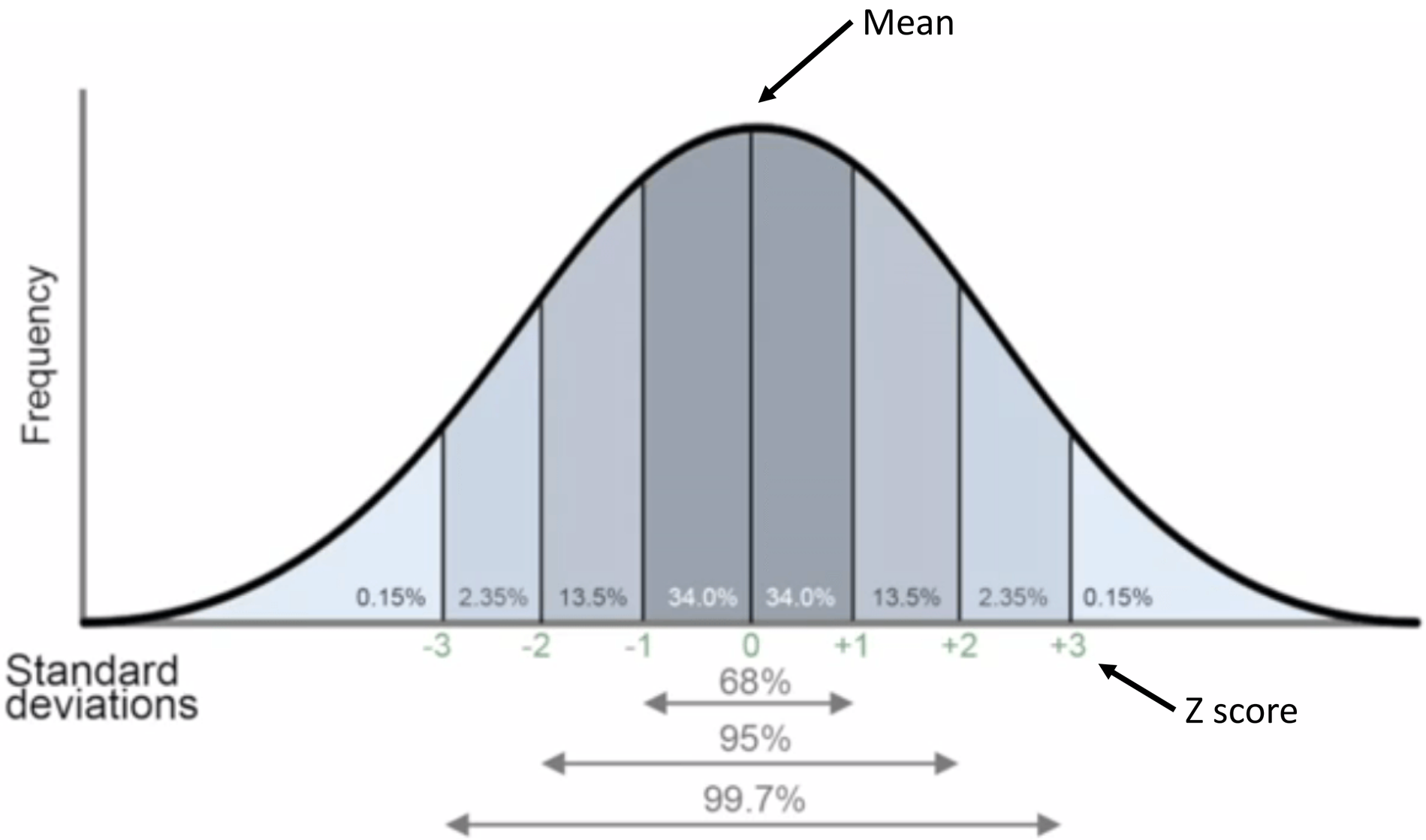
Definition: 3.8 *Standard coordinates*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. We represent these data items in standard coordinates by computing

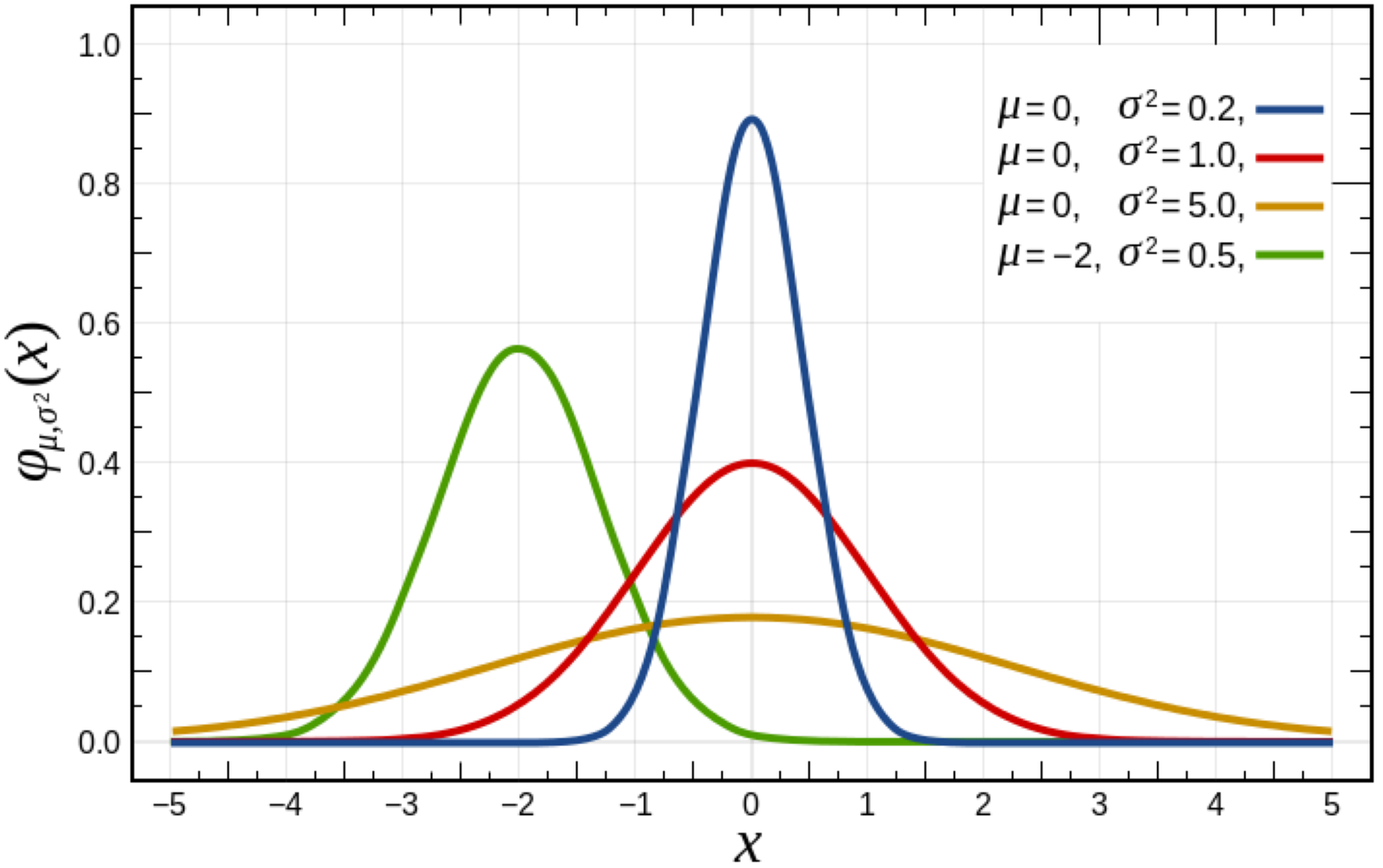
$$\hat{x}_i = \frac{(x_i - \text{mean}(\{x\}))}{\text{std}(x)}.$$

We write $\{\hat{x}\}$ for a dataset that happens to be in standard coordinates.

Number of standard deviations a point is away from mean

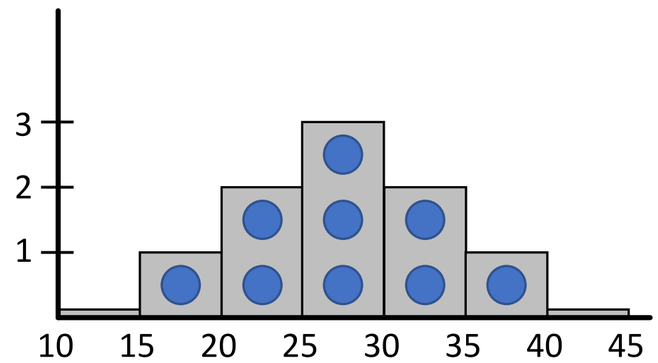

Mean
Frequency
0.15%
2.35%
13.5%
34.0%
34.0%
13.5%
2.35%
0.15%
-3
-2
-1
0
+1
+2
+3
Standard deviations
68%
95%
99.7%
Z score

Normal Distribution



An Example: Histogram

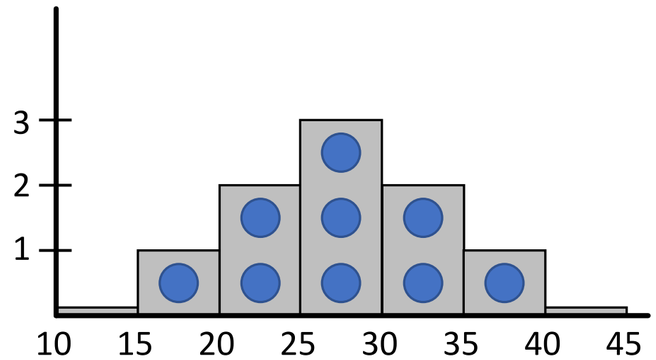
16
27
29
31
26
22
32
36
24

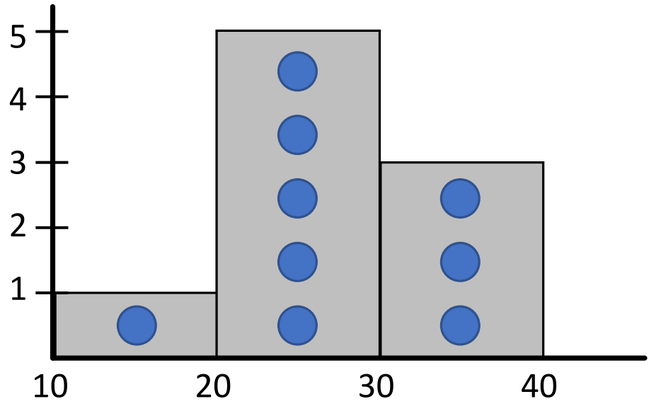


Mean (Average) = 27
Standard Deviation = 6

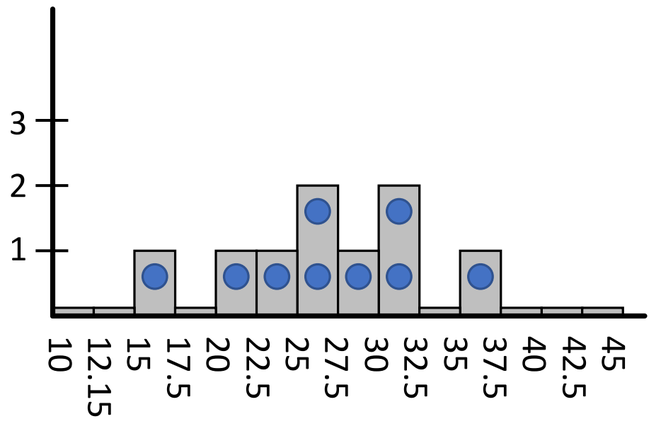
An Example: Histogram Resolution

16
27
29
31
26
22
32
36
24

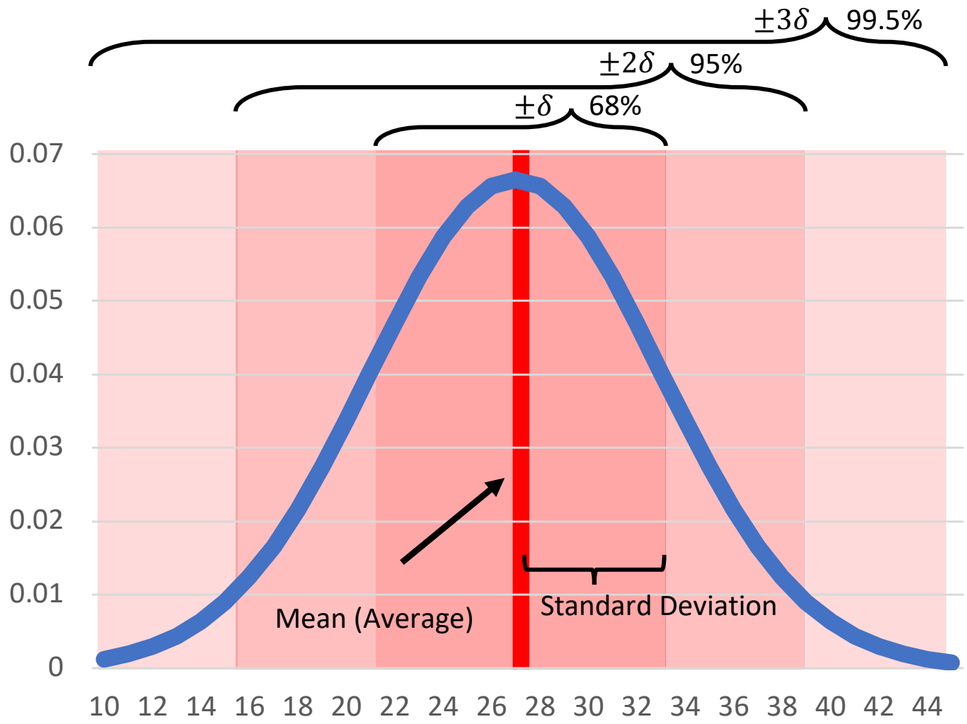


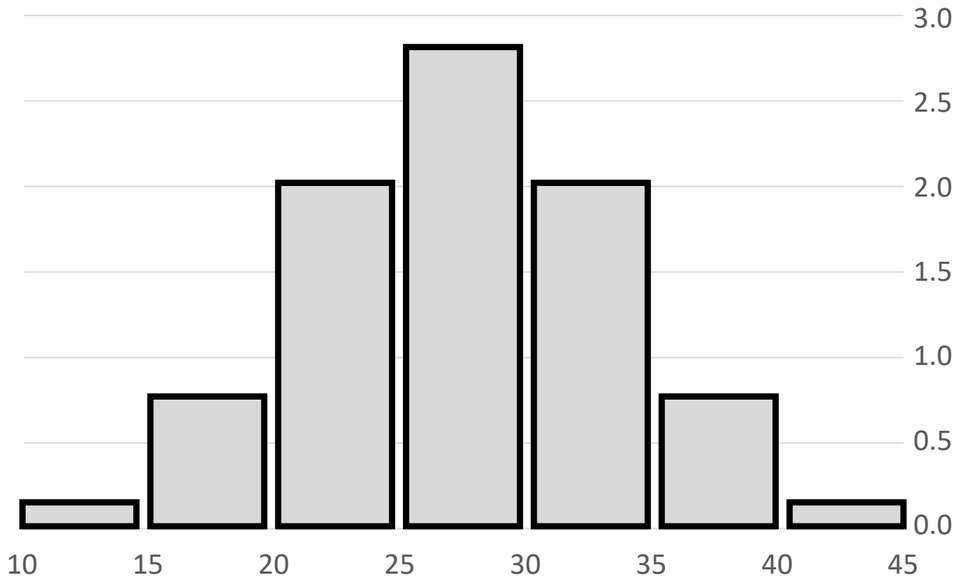


Mean (Average) = 27
Standard Deviation = 6

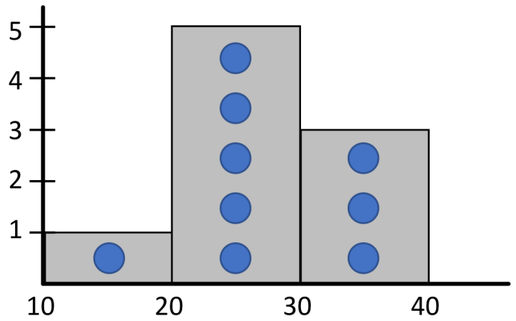


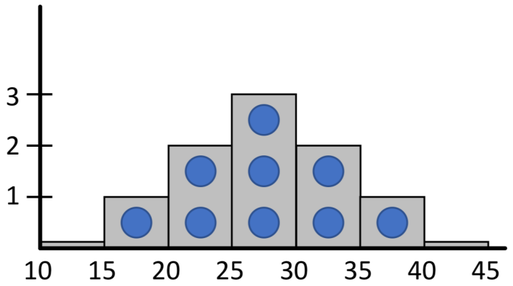
An Example: Statistical Distribution

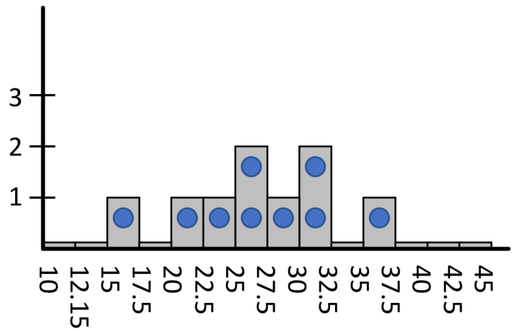


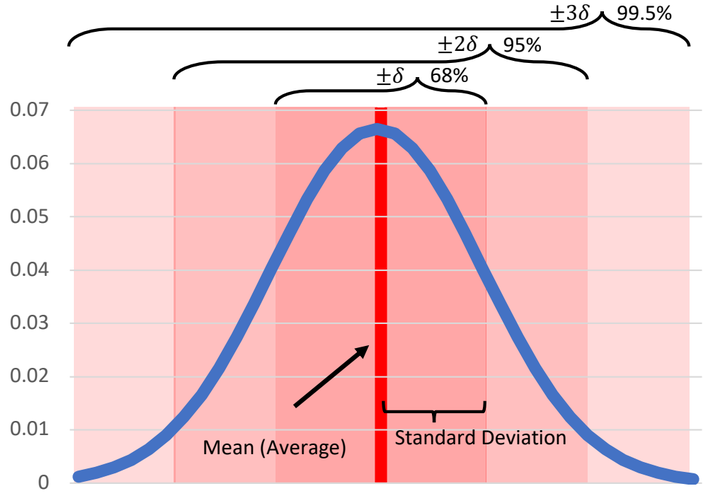


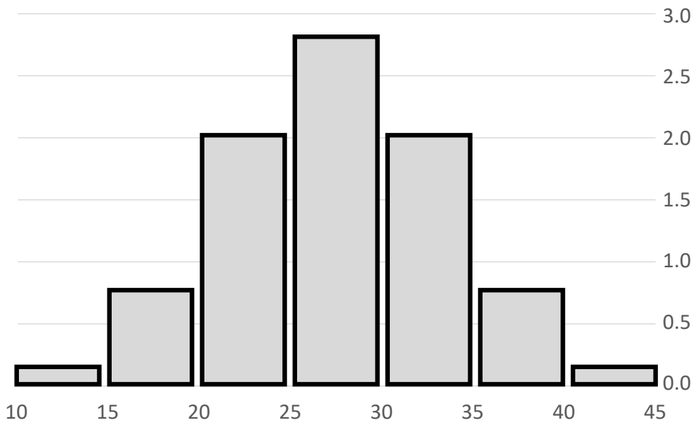
An Example: Comparing Histogram & Distribution





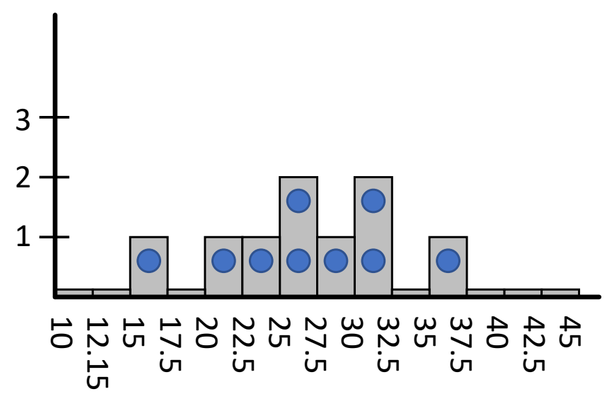


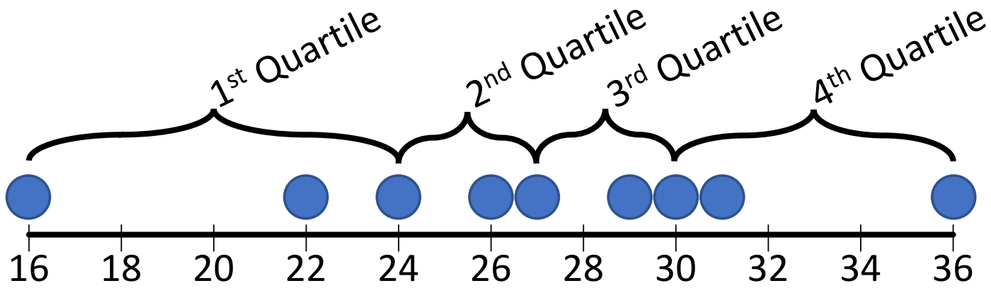




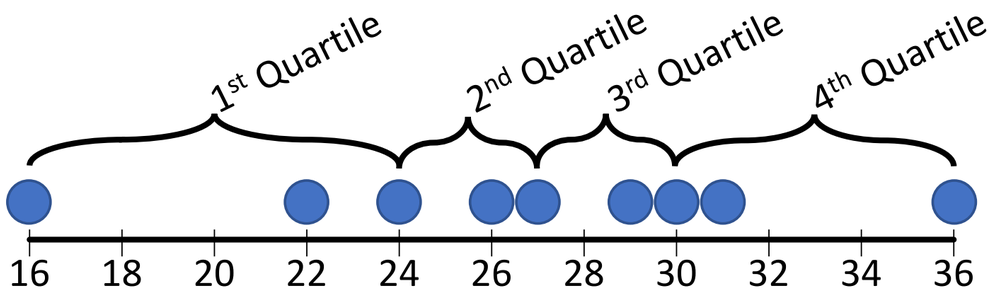
An Example: Comparing Histogram & Distribution

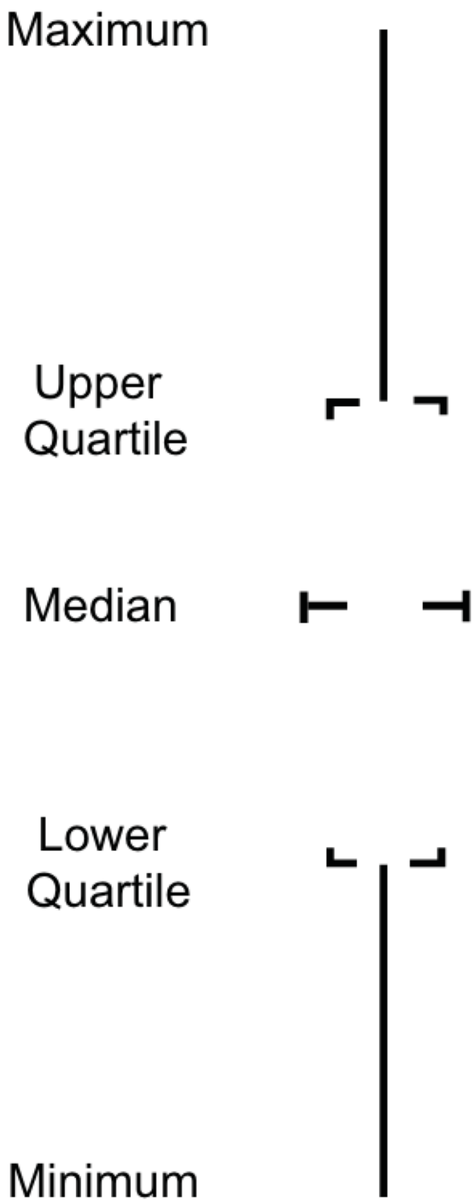
16
27
29
31
26
22
32
36
24





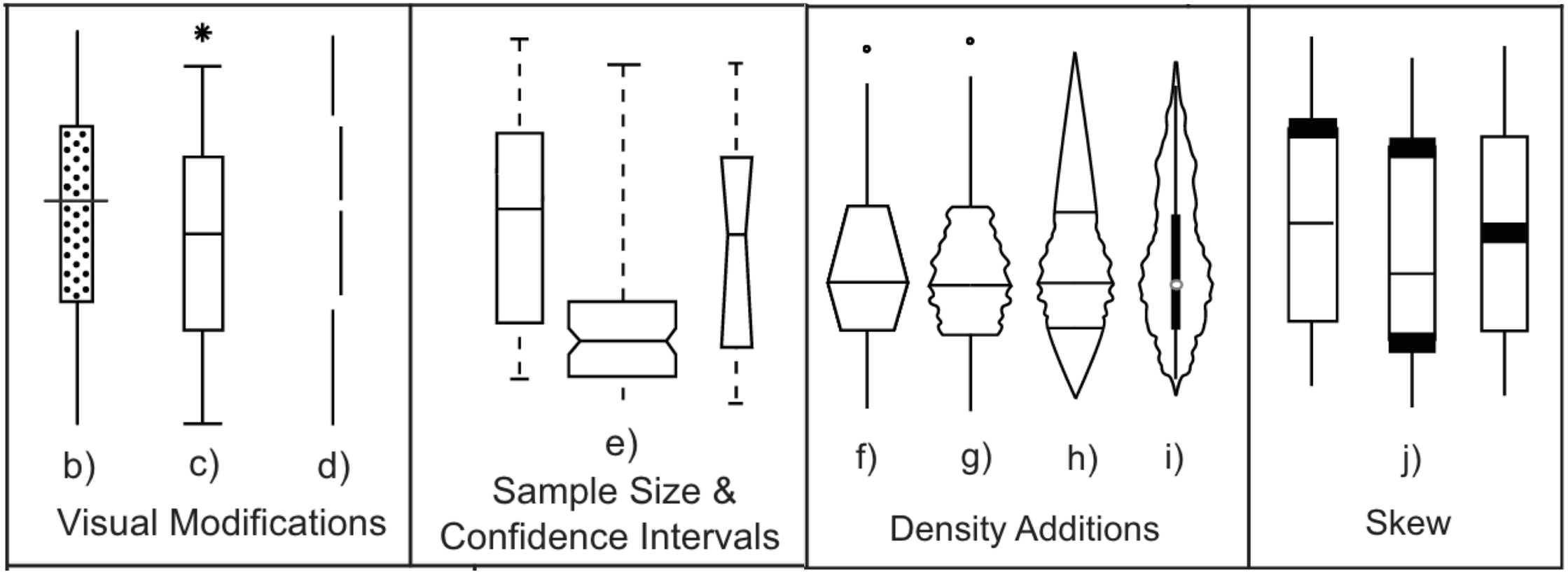
BOXPLOT





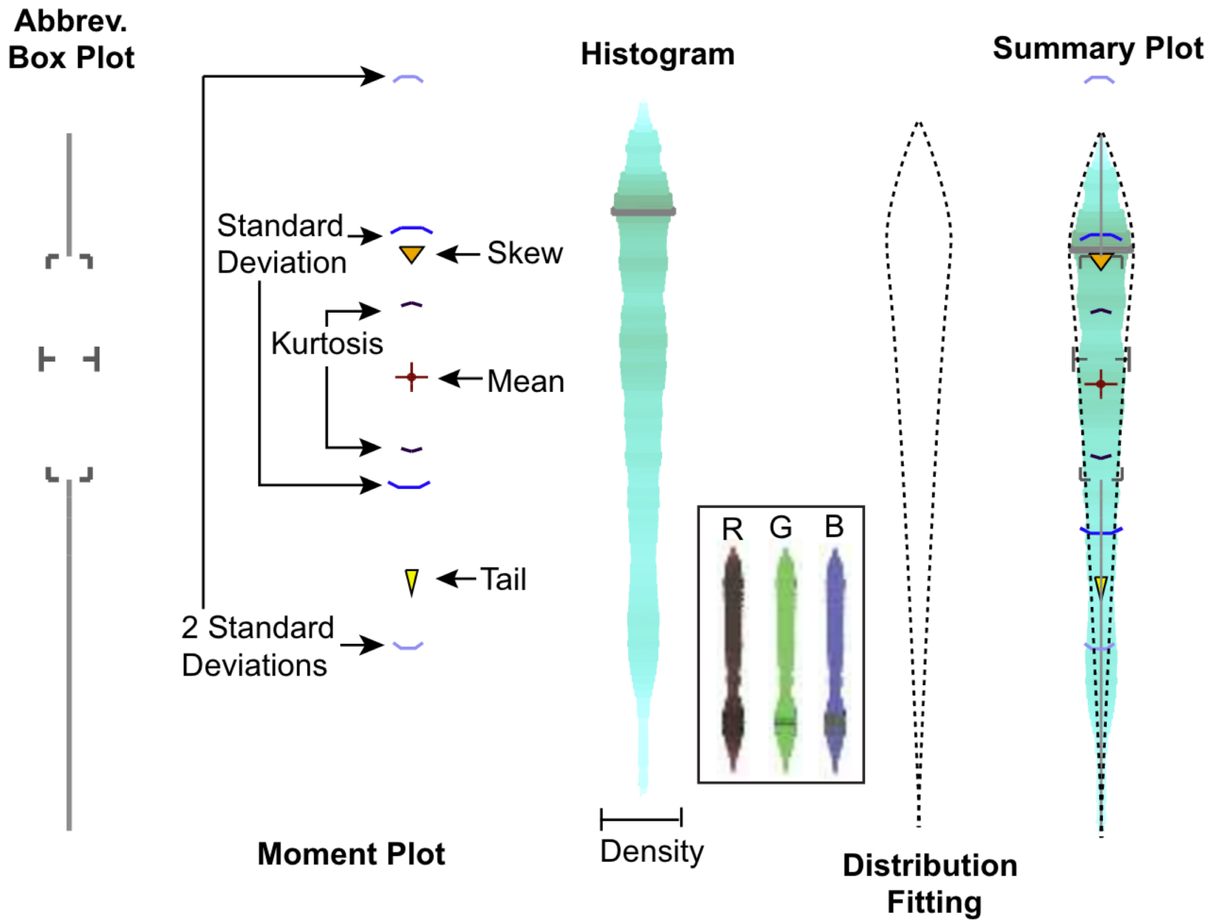
POTTER ET AL., 2010

BOXPLOTS



POTTER ET AL., 2010

BOXPLOTS



Given a data set $\{x_i\}_{i=1}^N$, we define the following quantities:

kth Central Moments:	$\mu_k \simeq \frac{1}{N}\sum_{i=1}^N (x_i - \mu_1)^k$
Mean:	$\mu_1 \simeq \frac{1}{N}\sum_{i=1}^N x_i$
Variance:	$\mu_2 \simeq \frac{1}{N}\sum_{i=1}^N (x_i - \mu_1)^2$
Standard Deviation:	$\sigma = \sqrt{\mu_2}$
Skew:	$\gamma = \frac{\mu_3}{\sigma^3}$
Kurtosis:	$\kappa = \frac{\mu_4}{\sigma^4}$
Excess Kurtosis:	$\kappa_e = \kappa - 3$
Tailing:	$\tau = \frac{\mu_5}{\sigma^5}$

where N is the number of data samples.

POTTER ET AL., 2010

Problem #2:

We have too many attributes to show

Pearson Correlation Coefficient

A measure of the linearity between 2 sets

$$\rho_{X,Y} = \frac{\mathbf{cov}(X,Y)}{\sigma_X \sigma_Y}$$

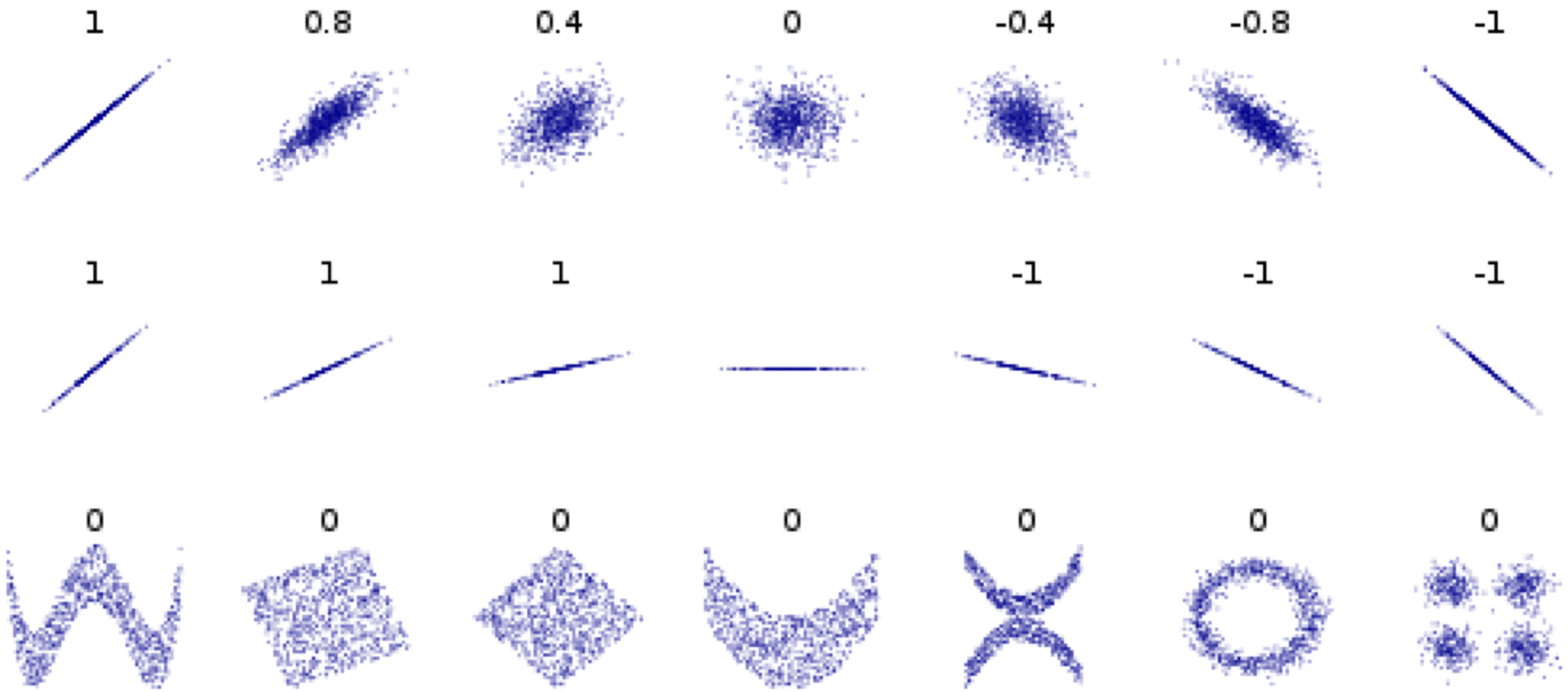
where:

- **cov** is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

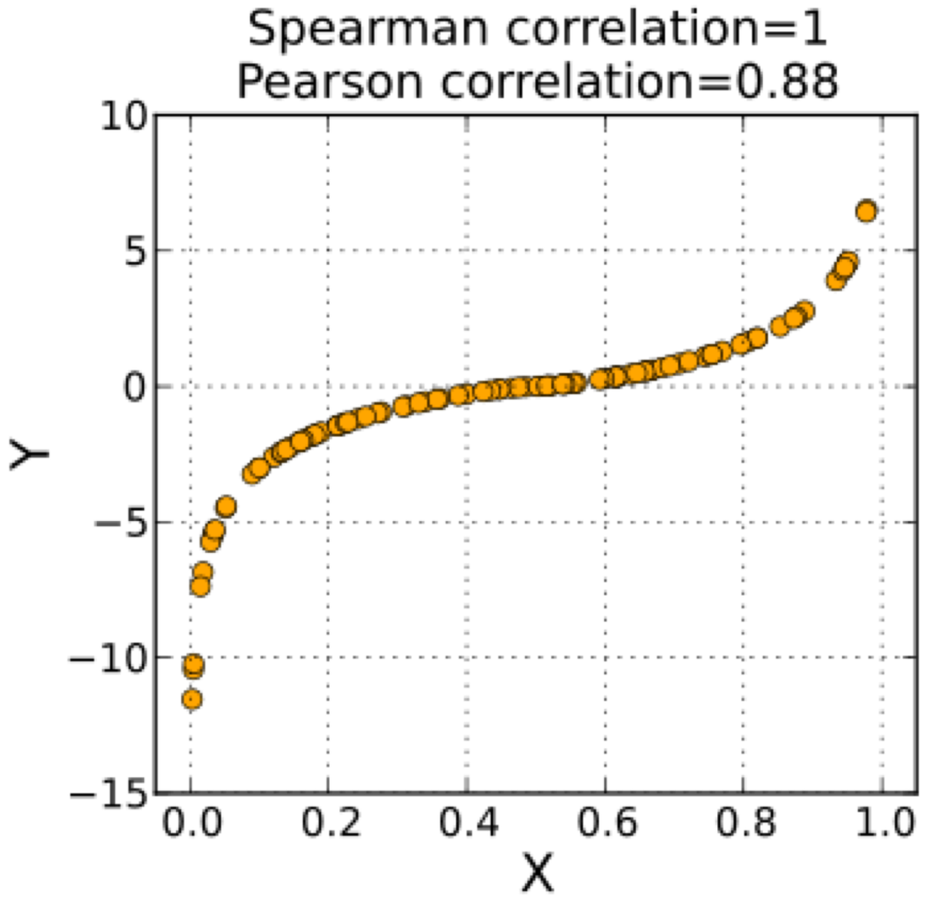
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n, x_i, y_i are defined as above
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

Spearman Rank Correlation



Spearman Rank Correlation

sort(X) and sort(Y)

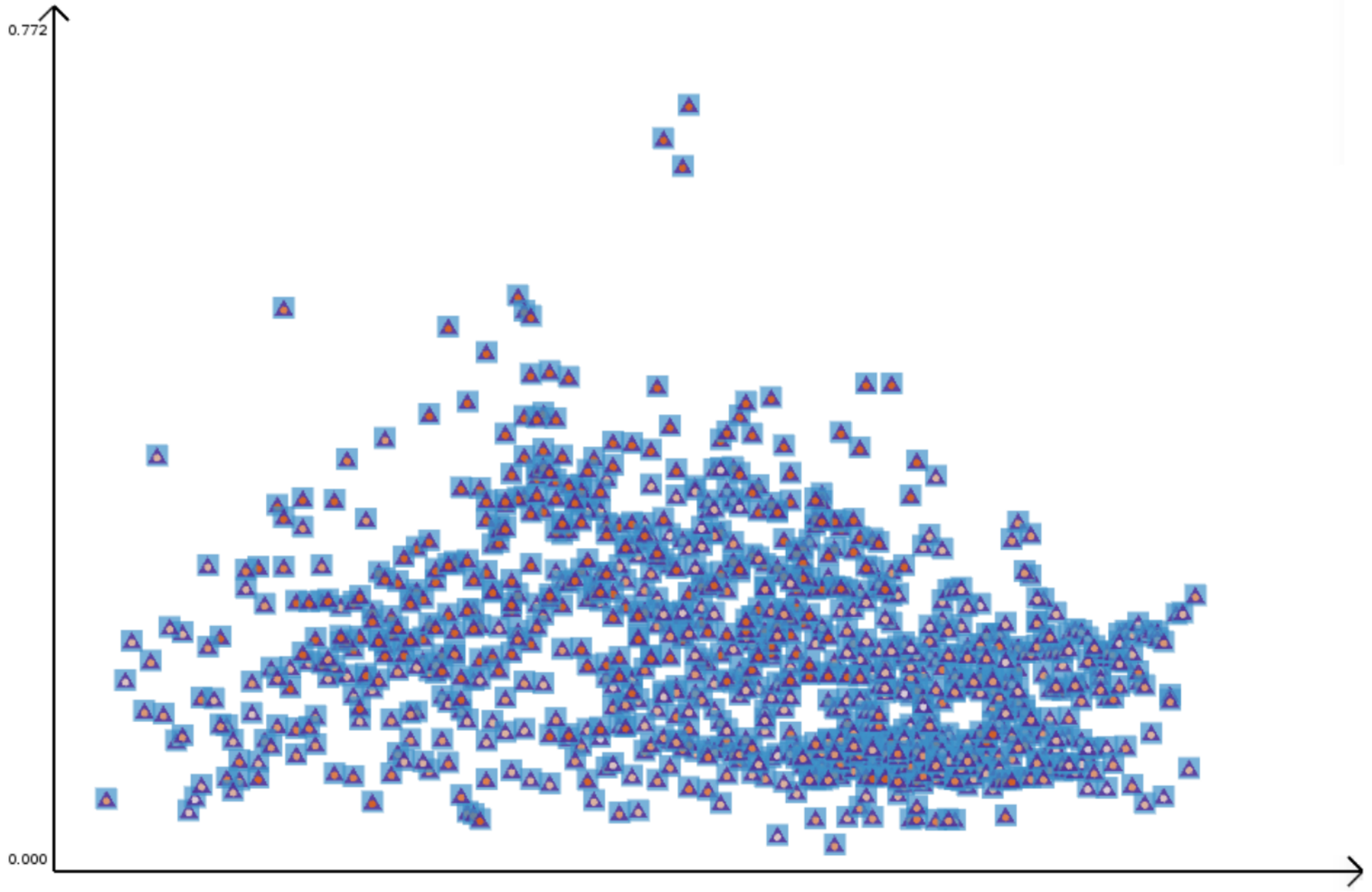
assign X'/Y' rank in sorted list

Calculate PCC(X', Y')

Spearman Rank Correlation

IQ, (X)	Hours of TV per week, (Y)	rank (X')	rank (Y')
86	0	1	1
97	20	2	6
99	28	3	8
100	27	4	7
101	50	5	10
103	29	6	9
106	7	7	3
110	17	8	5
112	6	9	2
113	12	10	4

MANY ATTRIBUTES MULTIPLE CORRELATION

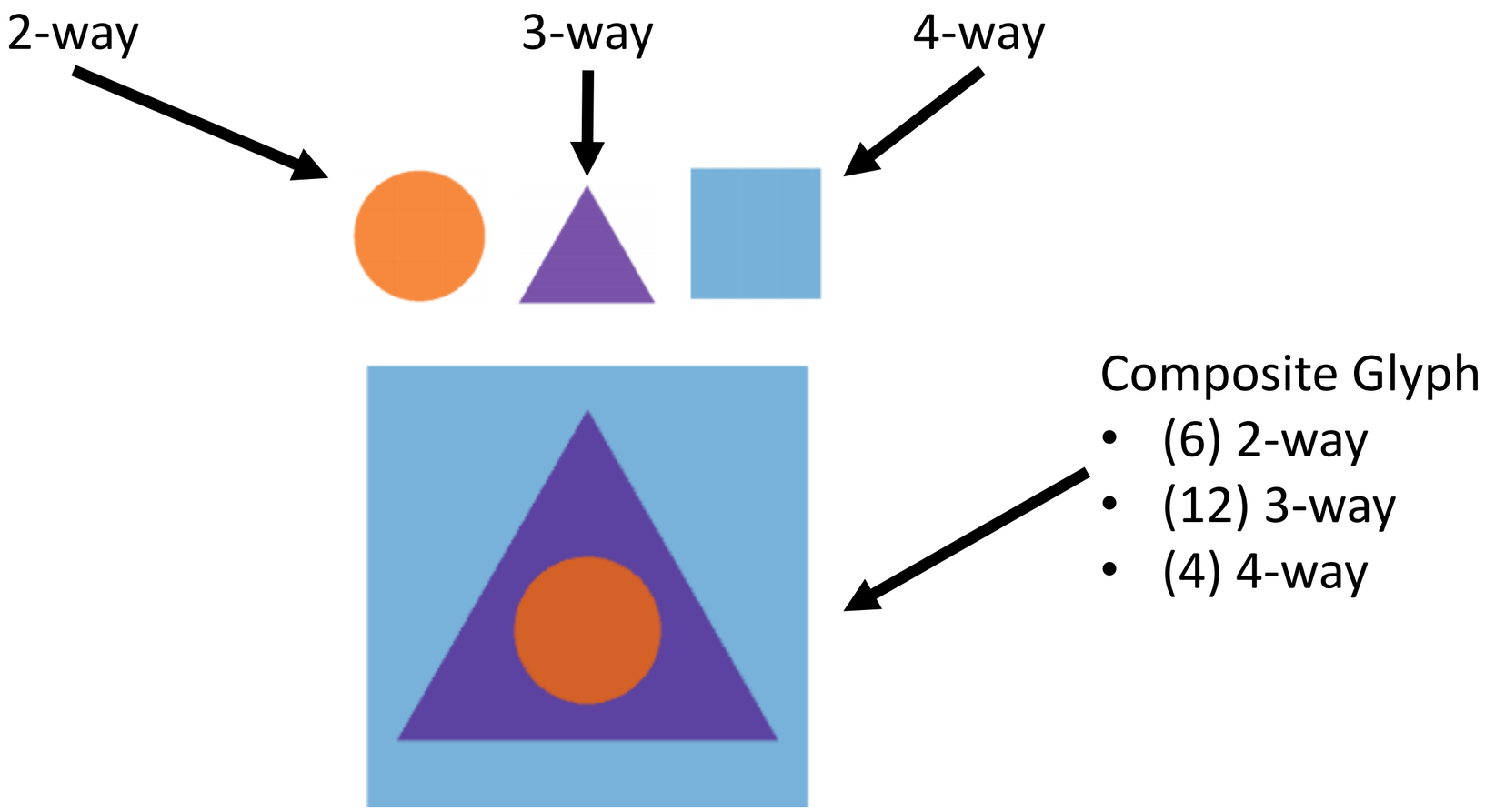


Multiple Correlation

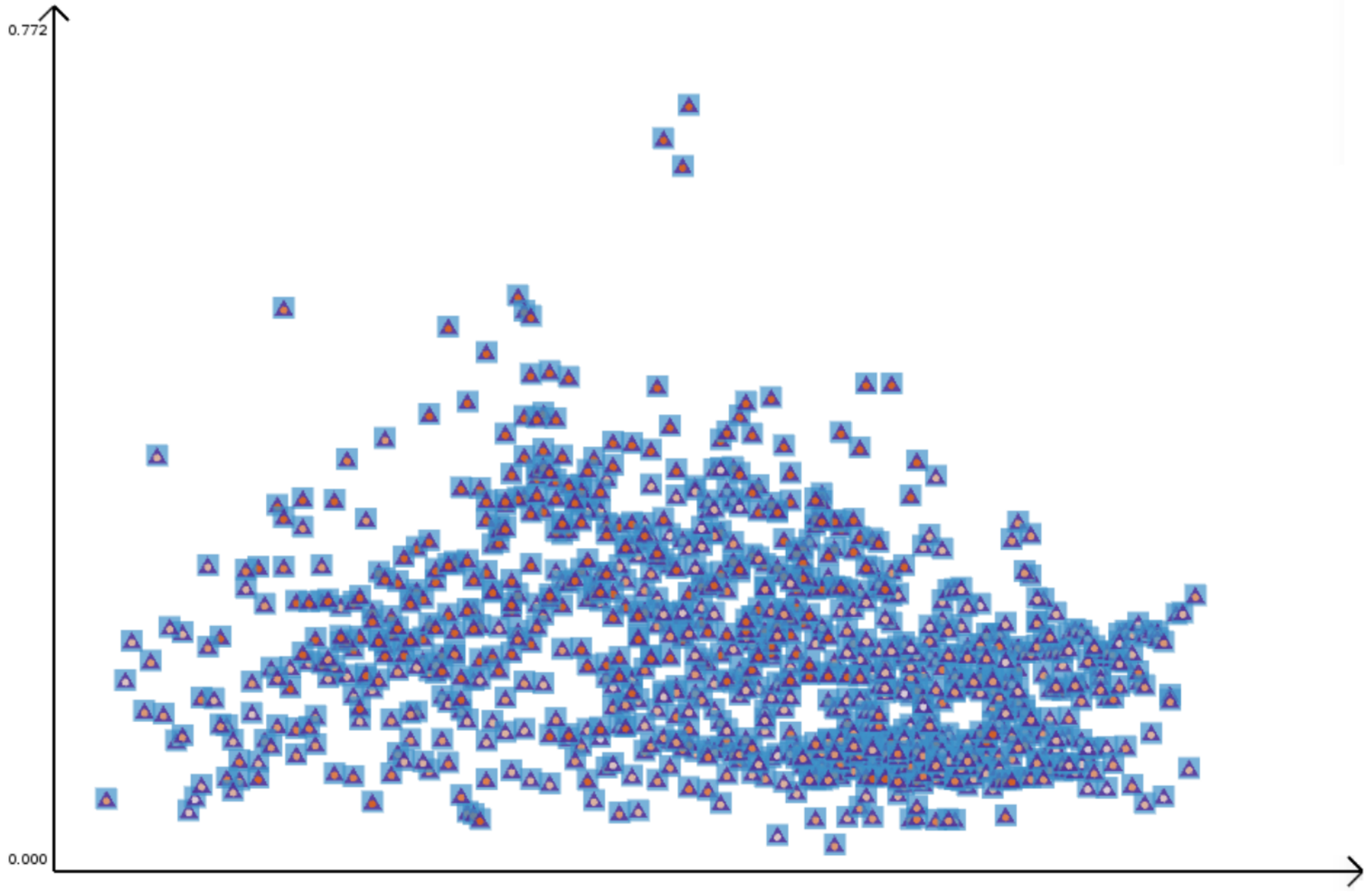
$$R^2 = \mathbf{c}^\top R_{xx}^{-1}\, \mathbf{c},$$

$$R_{xx} = \begin{pmatrix} r_{x_1 x_1} & r_{x_1 x_2} & \cdots & r_{x_1 x_N} \\ r_{x_2 x_1} & \ddots & & \vdots \\ \vdots & & \ddots & \\ r_{x_N x_1} & \cdots & & r_{x_N x_N} \end{pmatrix}.$$

MULTIPLE CORRELATION



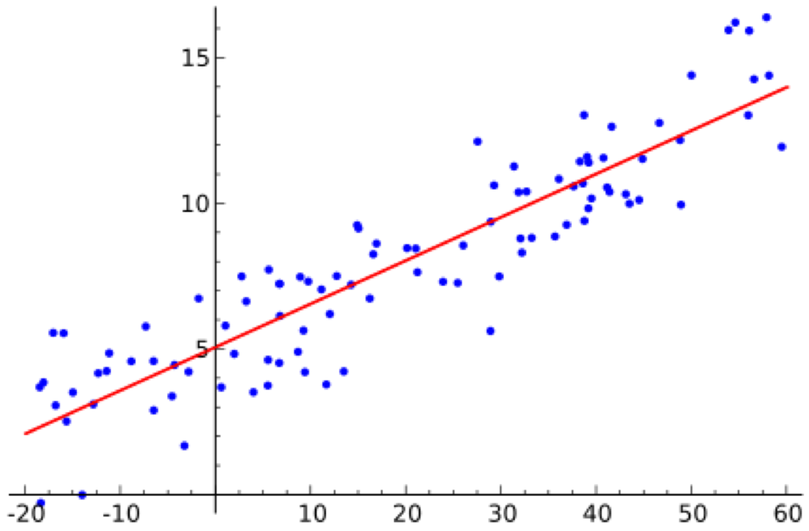
MANY ATTRIBUTES MULTIPLE CORRELATION



Regression: Fitting a Model to Data

Given: $y_i = \alpha + \beta x_i + \varepsilon_i$

Find α and β that minimize ε_i in the linear least squares sense (i.e. $\Sigma \varepsilon_i^2$)

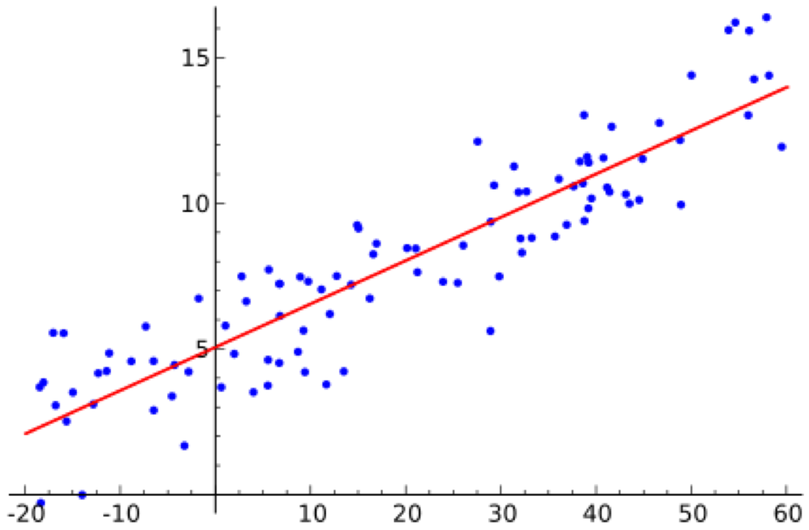


Regression: Fitting a Model to Data

Can be computed directly

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$



Problem #3

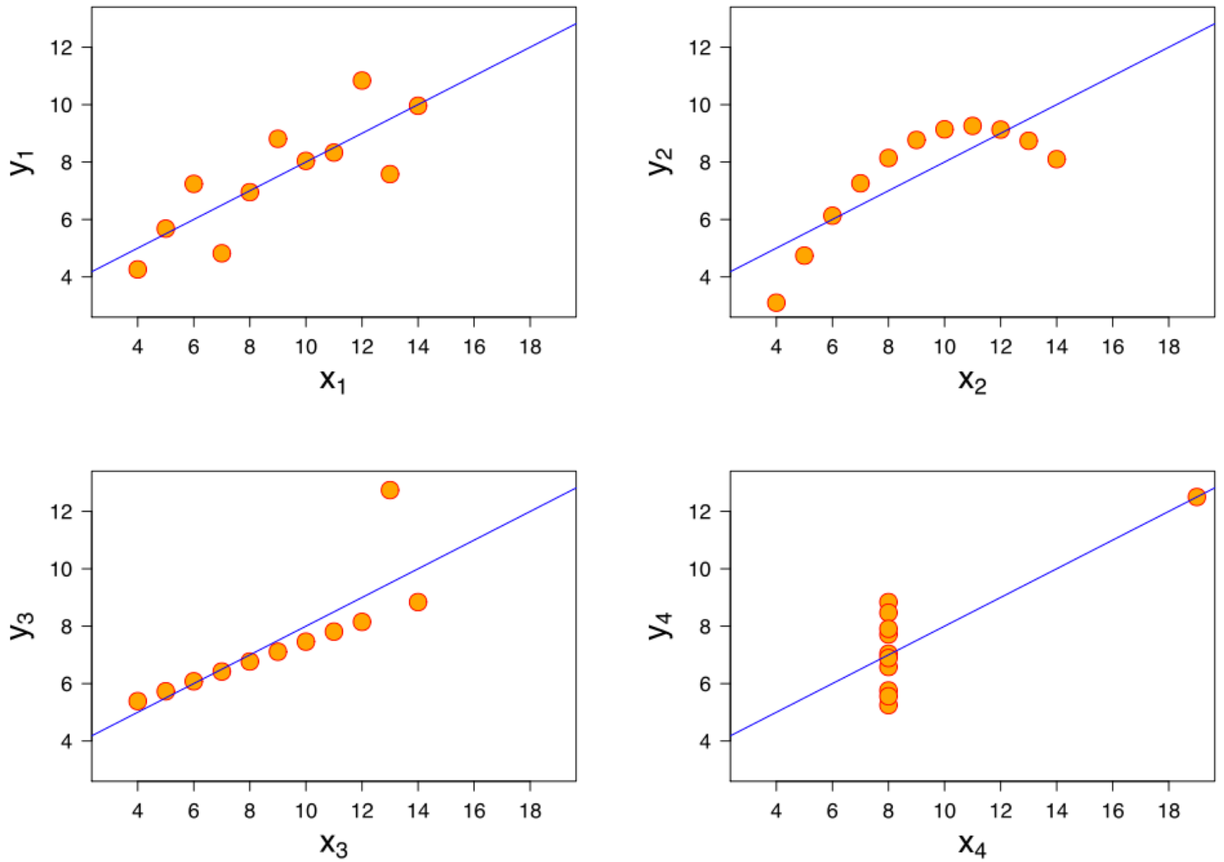
What is lost or misinterpreted…

So, what's the difference between Correlation and Regression and how does it impact our visualization?

KNOW THE SHAPES (INFORMATION) YOUR STATISTIC CAPTURES

Statistical Limitations

Anscombe's quartet

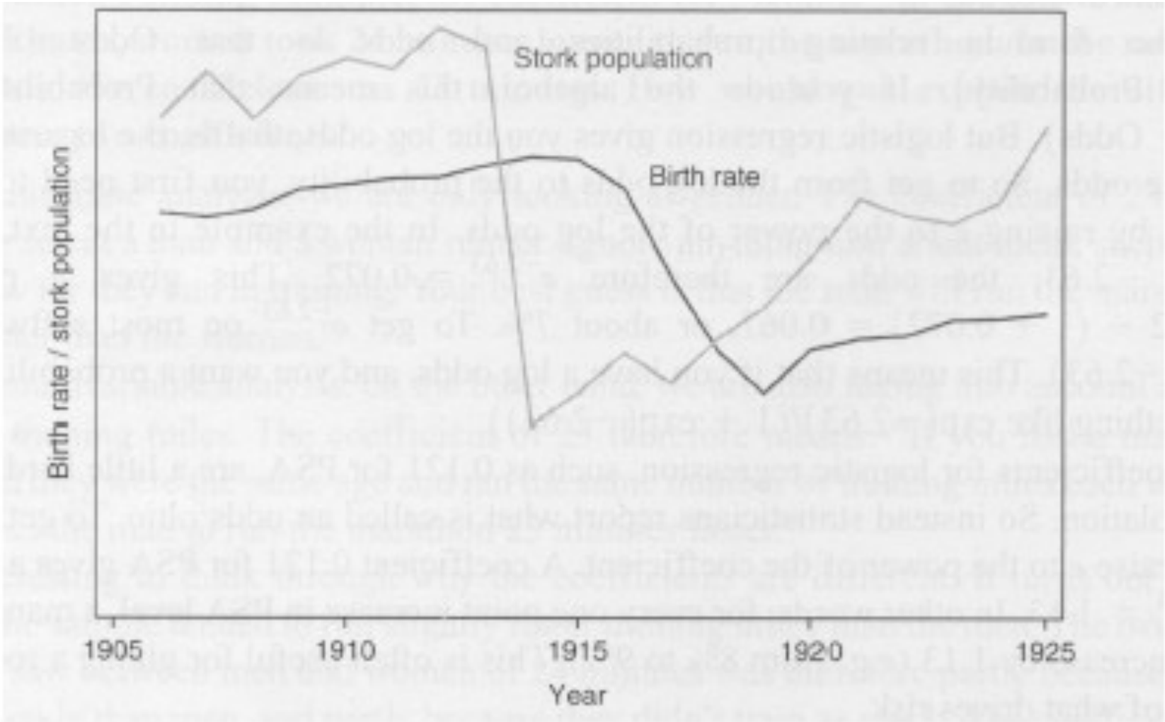


Statistical Limitations

Anscombe's quartet

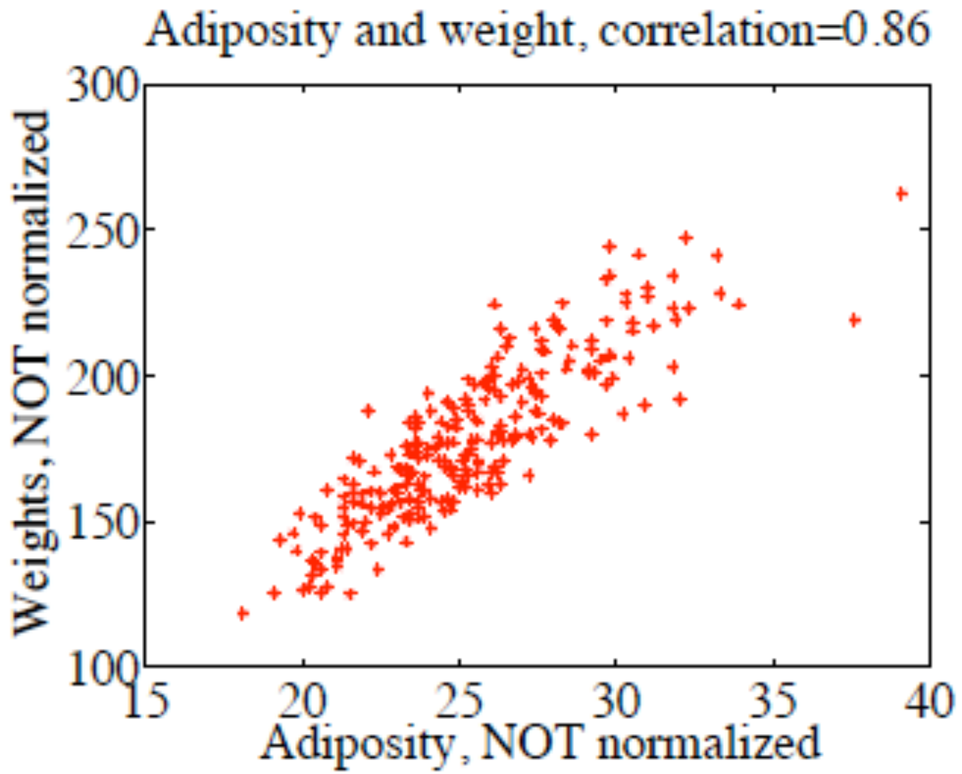
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	plus/minus 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	y = 3.00 + 0.500x	to 2 and 3 decimal places, respectively

CORRELATION != CAUSALITY



and foot size is positively correlated with reading ability, etc.

BUT CAN BE USED TO PREDICT

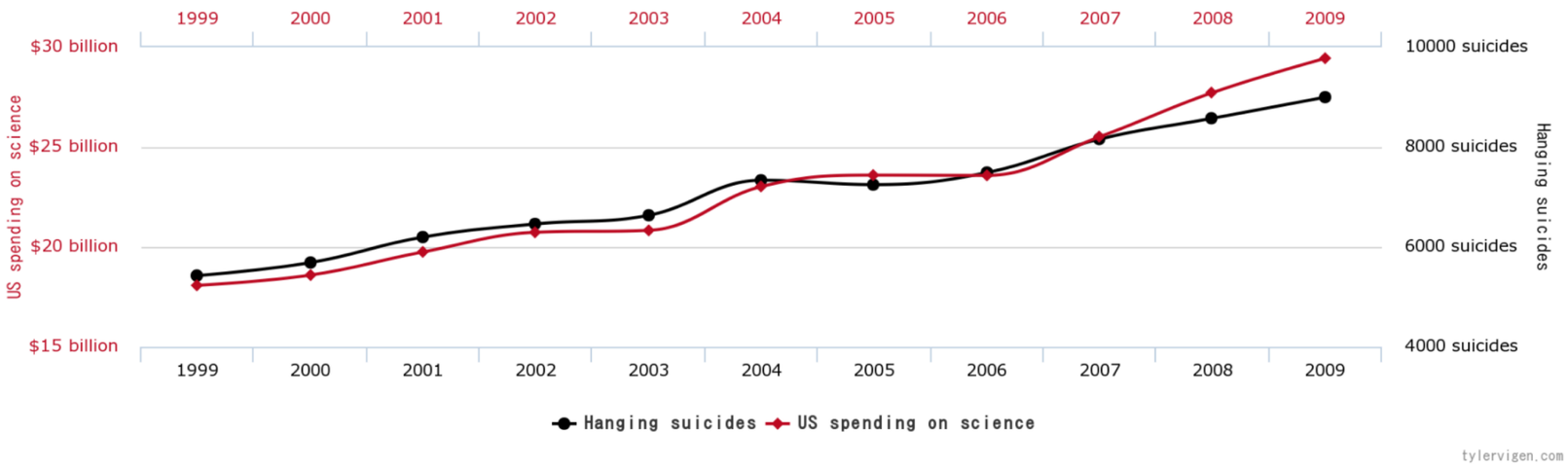


Spurious correlations

US spending on science, space, and technology

correlates with

Suicides by hanging, strangulation and suffocation

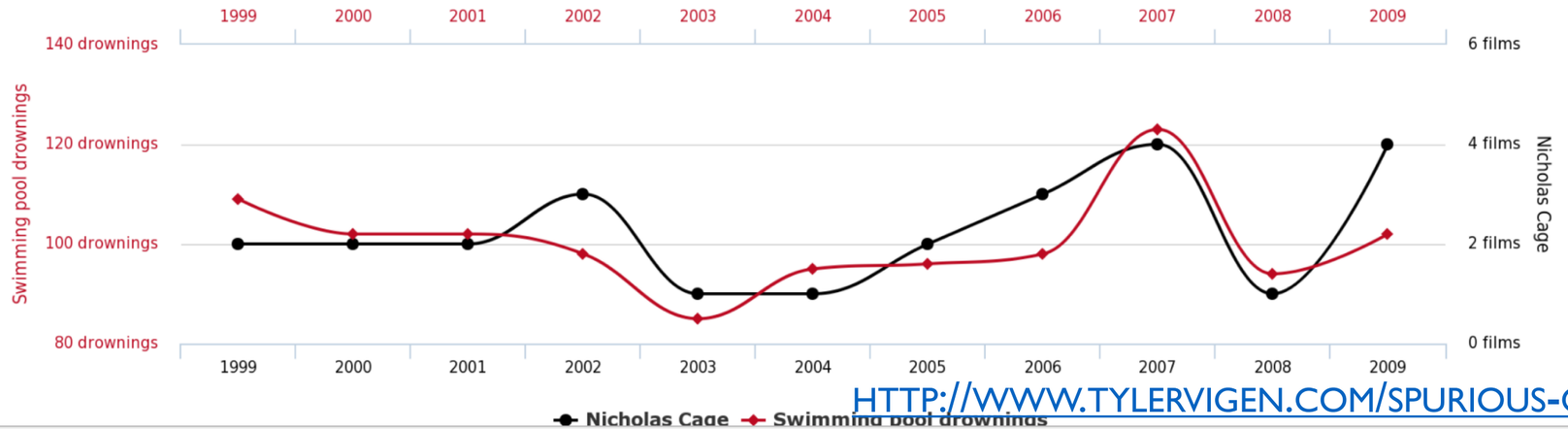


tylervigen.com

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in



HTTP://WWW.TYLERVIGEN.COM/SPURIOUS-CORRELATIONS

Statistical Vis Tools: R

It's free

It's easy to get pictures up and going

Many, many (almost too many) tools already available

Let's you work with tools *without* implementing them. However, you should UNDERSTAND them.

HTTP://STUDENTS.BROWN.EDU/SEEING-THEORY/INDEX.HTML