CS 571 - Data Visualization & Exploration

Filtering and Aggregation

Instructor: Hamza Elhamdadi

UMassAmherst

Upcoming Dates

Apr 25: Homework 4 (Due at 11:59pm Eastern)

Apr 24: Quiz 6 Released (all quizzes due May 2)

Apr 29: Final Group Activity

May 2:

- Homework 5 (Due at 11:59pm Eastern)
- Project Screencast Submission

May 12: Final Project Submission

Reducing Items and Attributes

Why Reduce?

Often, showing all of the data items and/or too many attributes will obscure the important features of the data

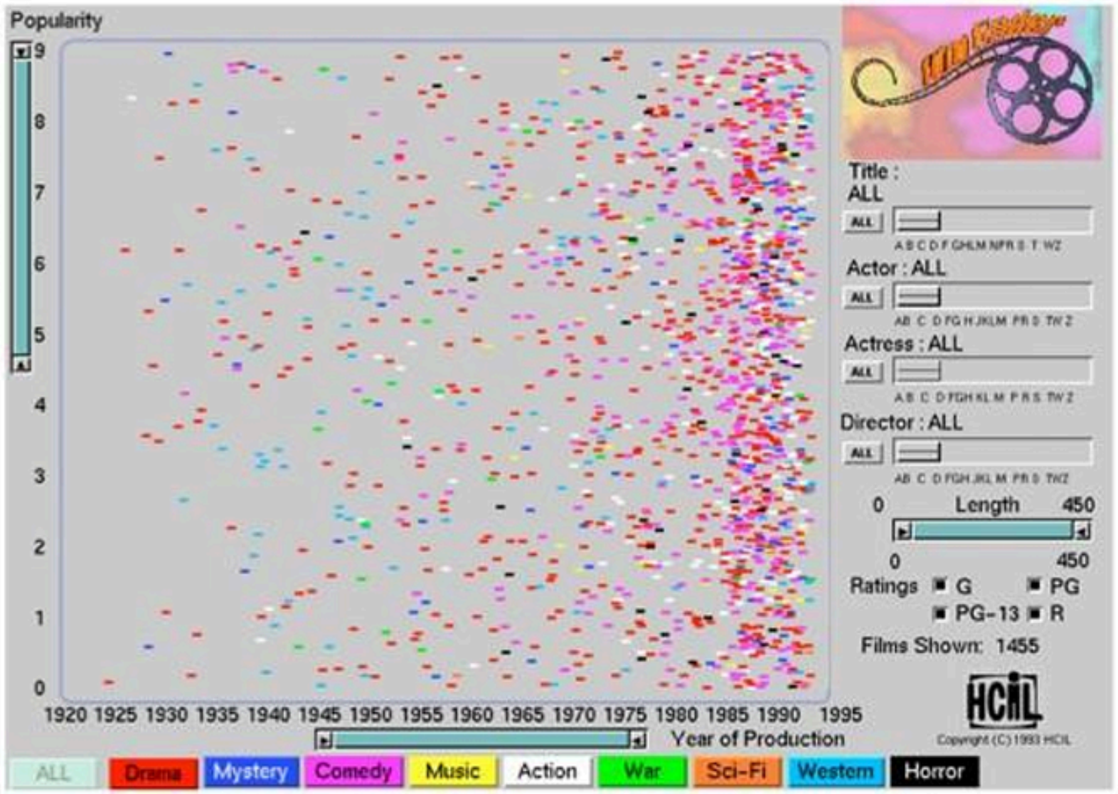
Filter

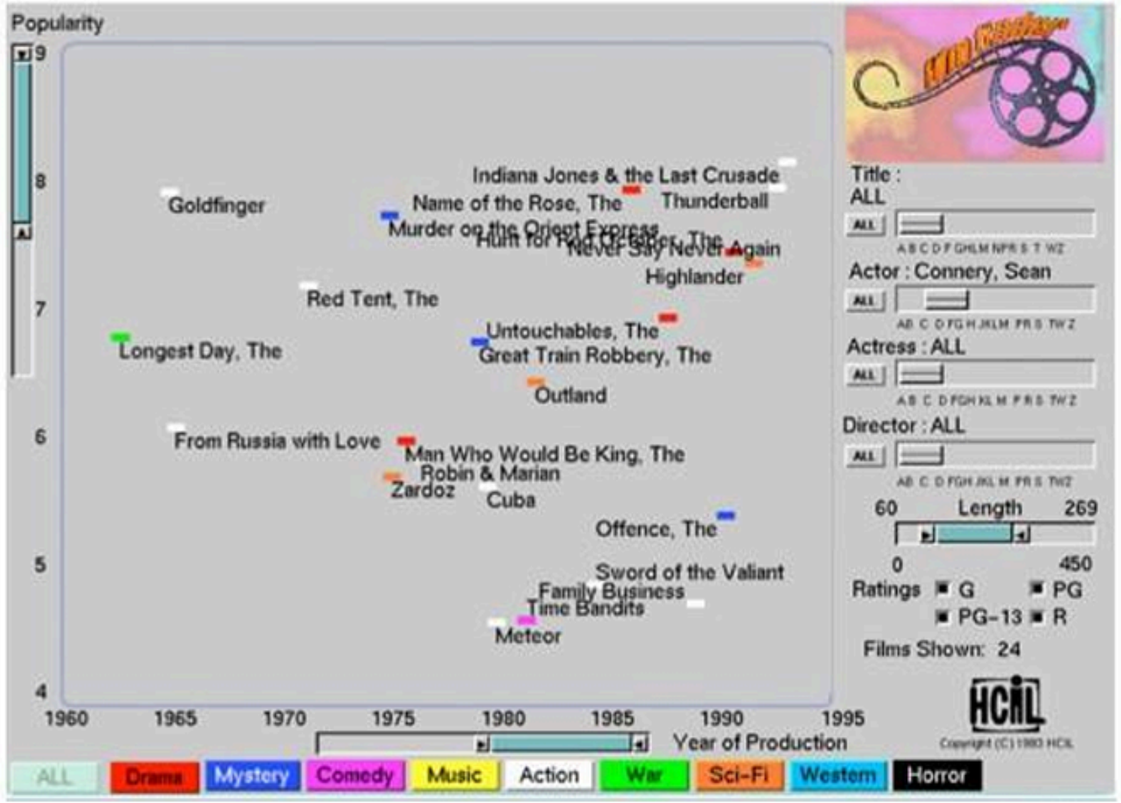
Filter

Elements are eliminated to support dynamic queries

- interaction + encoding
- user should see the immediate results of their actions

Item Filtering

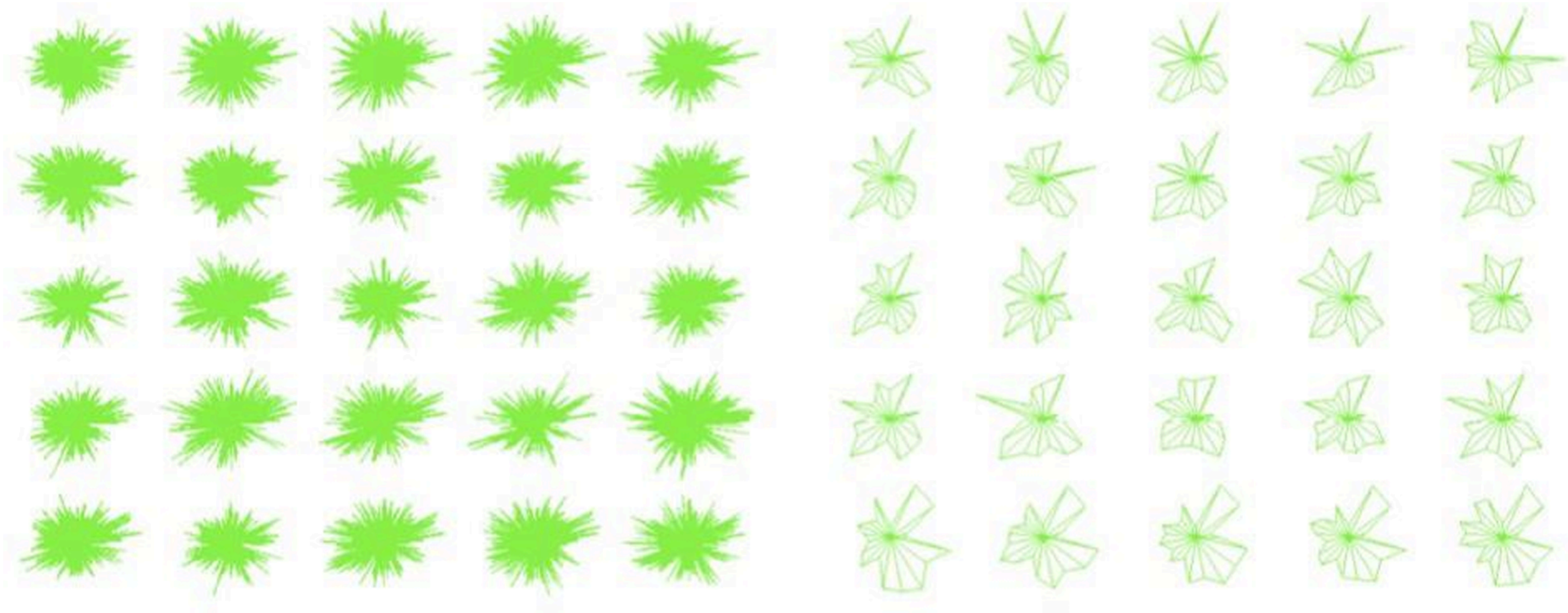




Item Filtering

https://archive.nytimes.com/www.nytimes.com/interactive/dining/new-york-health-department-restaurant-ratings-map.html

Attribute Filtering

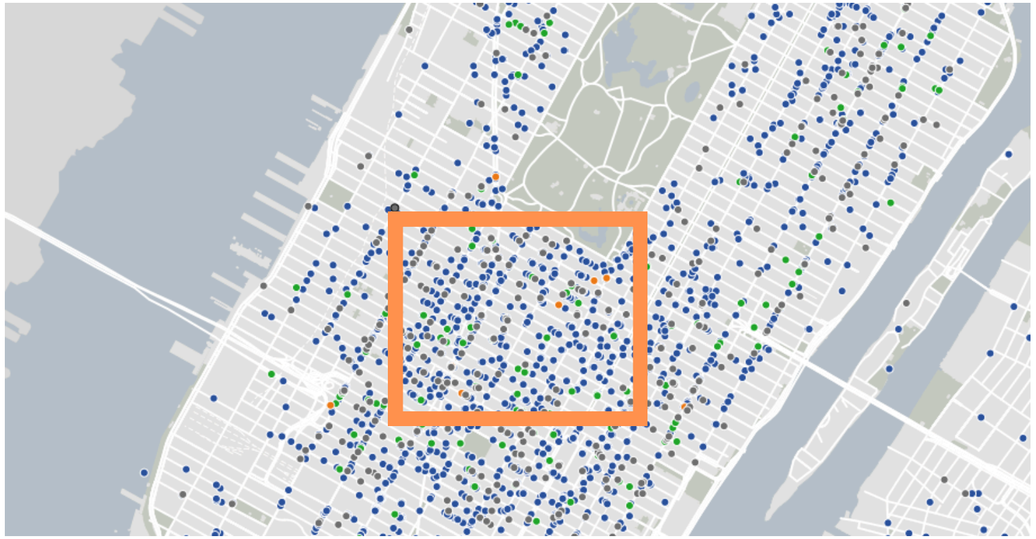


Filter Controls

2 Different Approaches:

- Widget-based filtering

- Visualization-based filtering



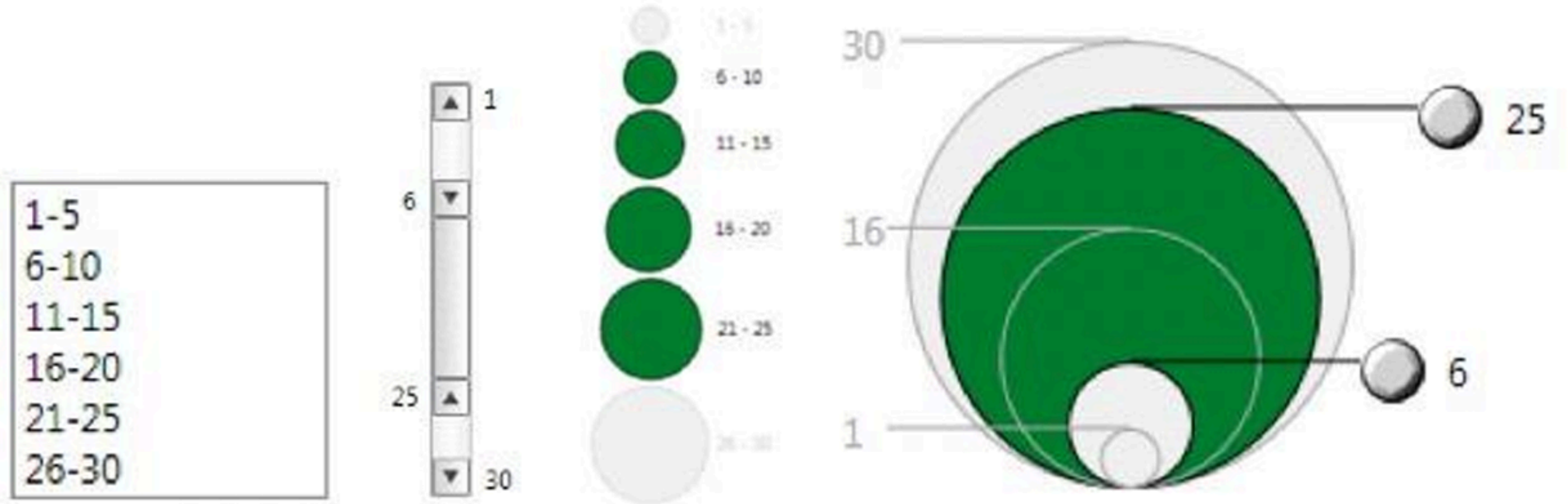
Filter Controls: Scented Widgets

Information Scent:

- gives the user a sense of the data through visual cues

Filter Controls: Interactive Legends

Controls that combine the visual aesthetic of static legends with the interaction mechanisms of widgets



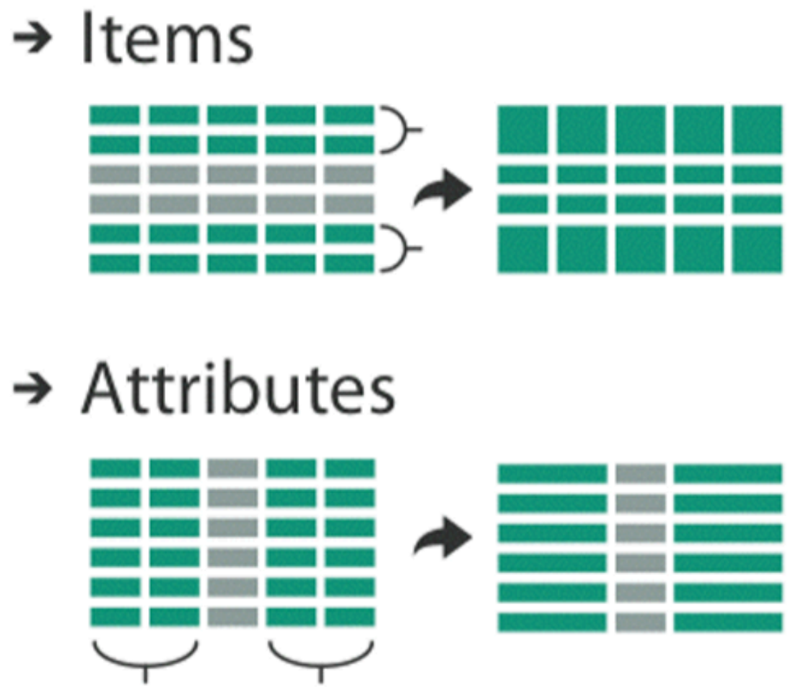
Aggregate

Aggregate

Group of elements is represented by a new derived element that stands in for the entire group

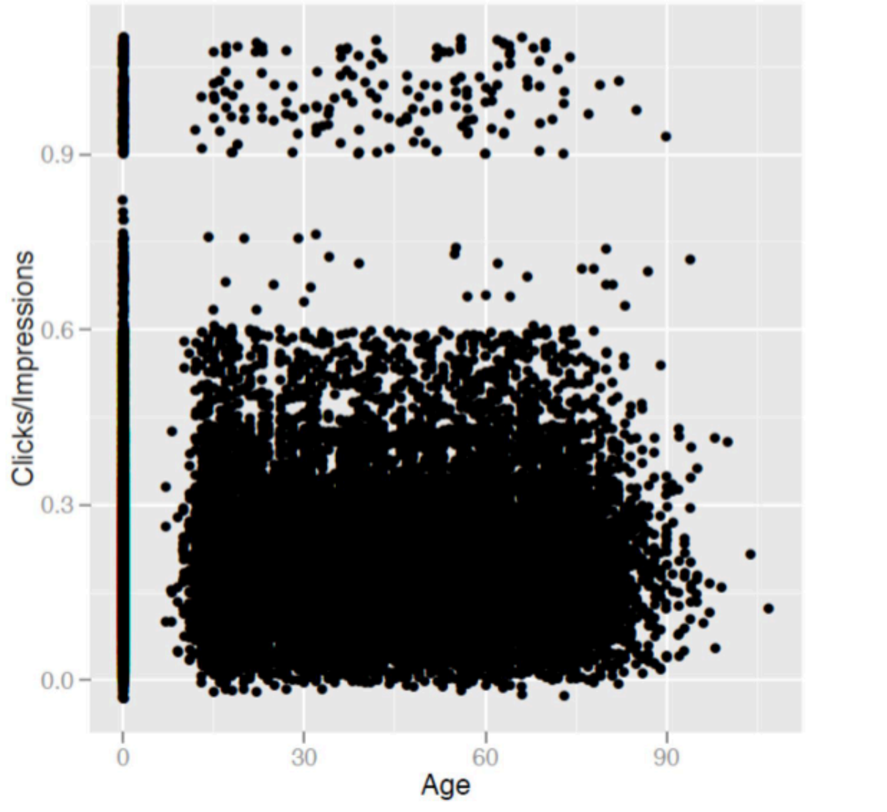
Many different ways to aggregate:

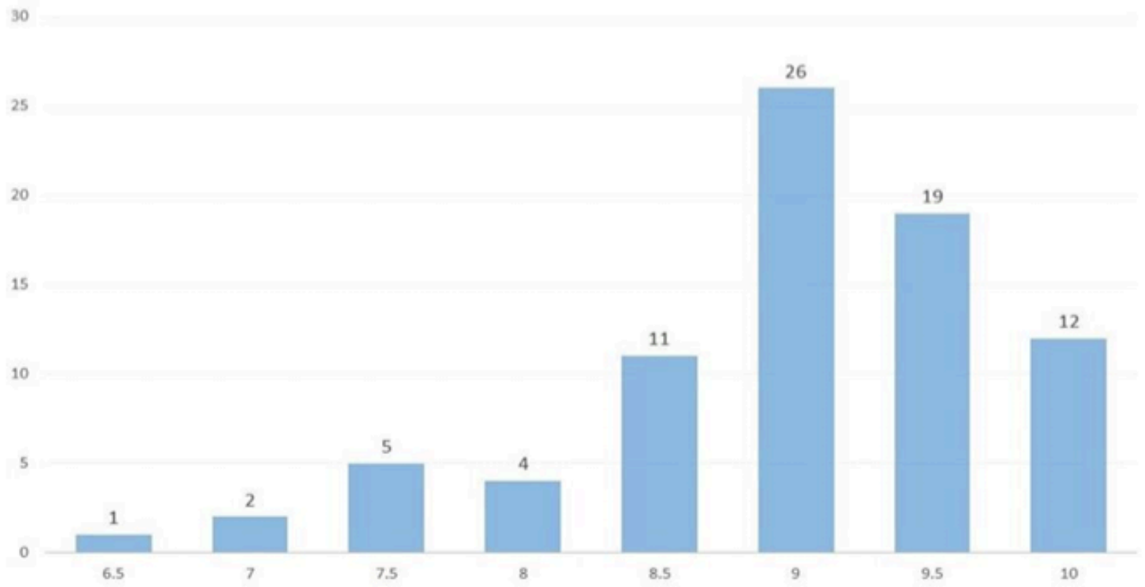
- Stats, Topology, ML, …



Problem 1: Aggregate Items

Too many items to show

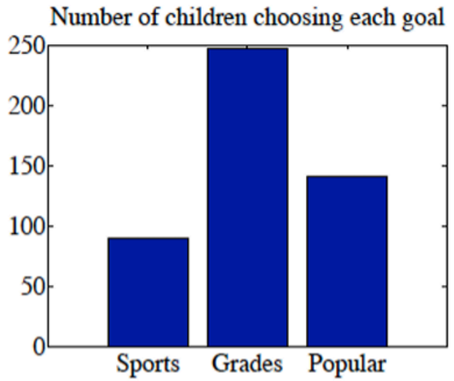


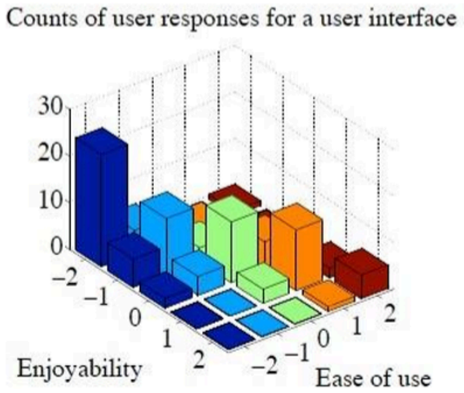


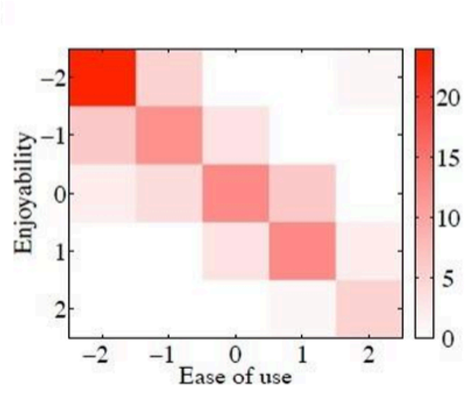
Problem 1: Aggregate Items

Histogram:

- a visualization that affords evaluating distribution of values





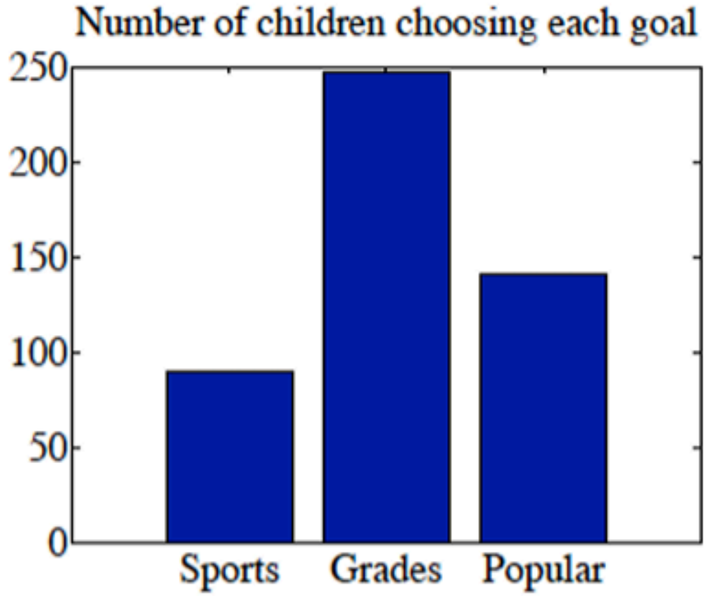


Problem 1: Aggregate Items

Histogram (**categorical**):

- simply count the occurrences of each category

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

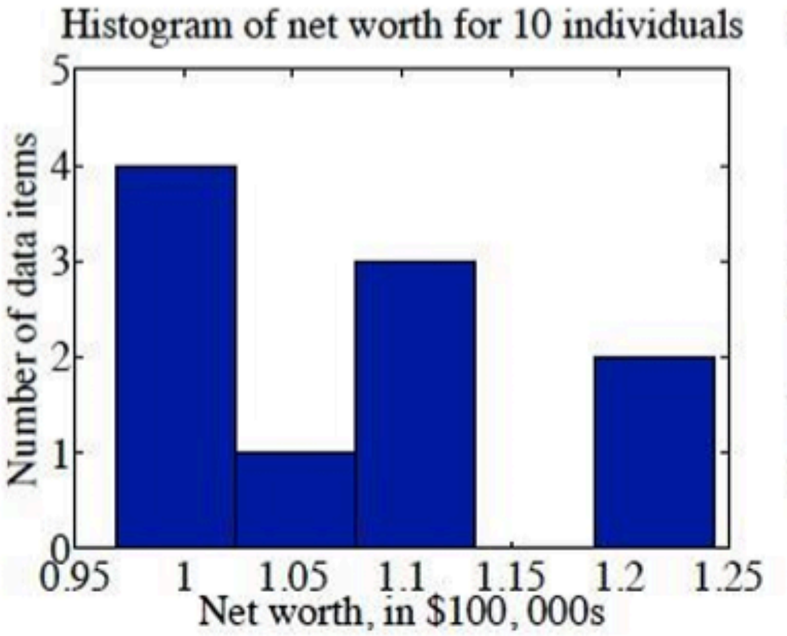


Problem 1: Aggregate Items

Histogram (**continuous**):

- **can't** simply count the occurrences of each category

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



Problem 1: Aggregate Items

Calculating a Continuous Histogram

$$\text{Given } \{x_0, \ldots, x_n\}$$

$$\text{Select } k \text{ bins}$$

$$bin_i = \lfloor k \times \frac{x_i - min(X)}{max(X) - min(X)} \rfloor$$

Problem 1: Aggregate Items

Calculating a Continuous Histogram

$$X = \{1, 2.5, 3, 4\}$$

$$k = 3$$

Problem 1: Aggregate Items

Calculating a Continuous Histogram

$$X = \{1, 2.5, 3, 4\}$$

$$k = 3$$

$$bin_i = \lfloor 3 \times \frac{1 - 1}{4 - 1} \rfloor = bin_0$$

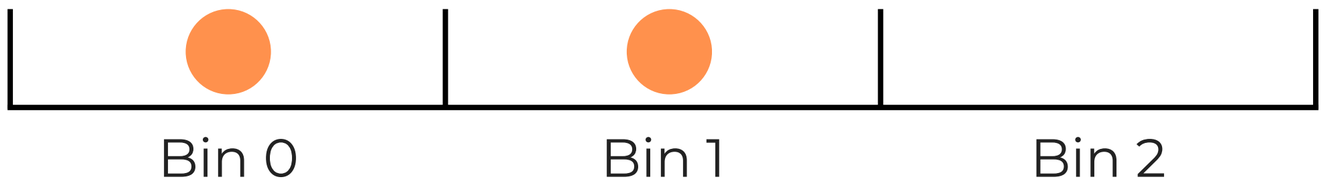
Problem 1: Aggregate Items

Calculating a Continuous Histogram

$$X = \{1, 2.5, 3, 4\}$$

$$k = 3$$

$$bin_i = \lfloor 3 \times \frac{2.5 - 1}{4 - 1} \rfloor = bin_1$$



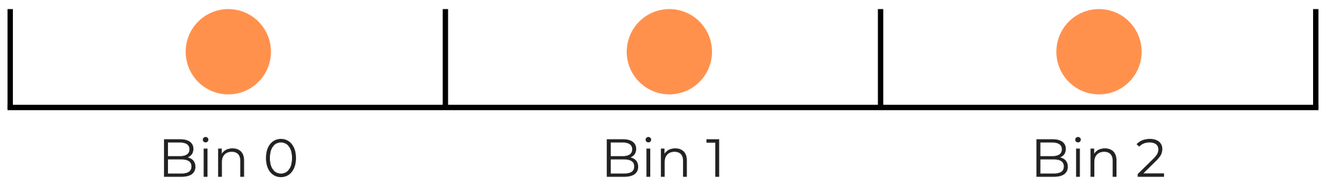
Problem 1: Aggregate Items

Calculating a Continuous Histogram

$$X = \{1, 2.5, 3, 4\}$$

$$k = 3$$

$$bin_i = \lfloor 3 \times \frac{3 - 1}{4 - 1} \rfloor = bin_2$$



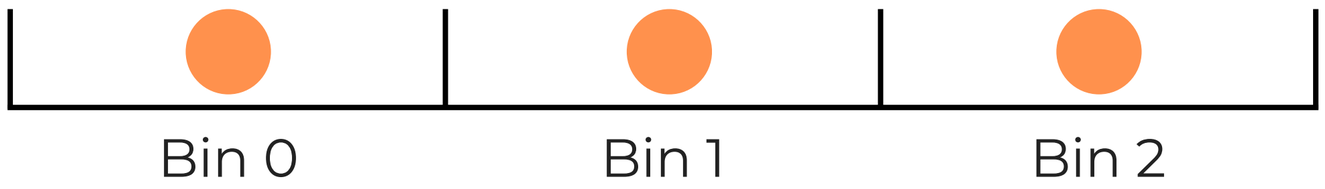
Problem 1: Aggregate Items

Calculating a Continuous Histogram

$$X = \{1, 2.5, 3, 4\}$$

$$k = 3$$

$$bin_i = \lfloor 3 \times \frac{4 - 1}{4 - 1} \rfloor = bin_3?$$



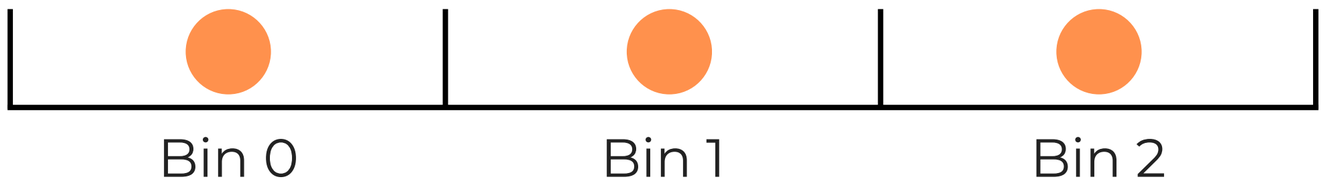
Problem 1: Aggregate Items

Calculating a Continuous Histogram

$$X = \{1, 2.5, 3, 4\}$$

$$k = 3$$

$$bin_i = min\{k - 1, \lfloor 3 \times \frac{4 - 1}{4 - 1} \rfloor\}$$



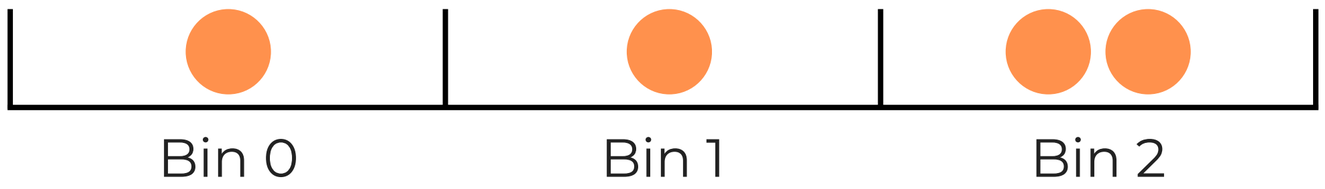
Problem 1: Aggregate Items

Calculating a Continuous Histogram

$$X = \{1, 2.5, 3, 4\}$$

$$k = 3$$

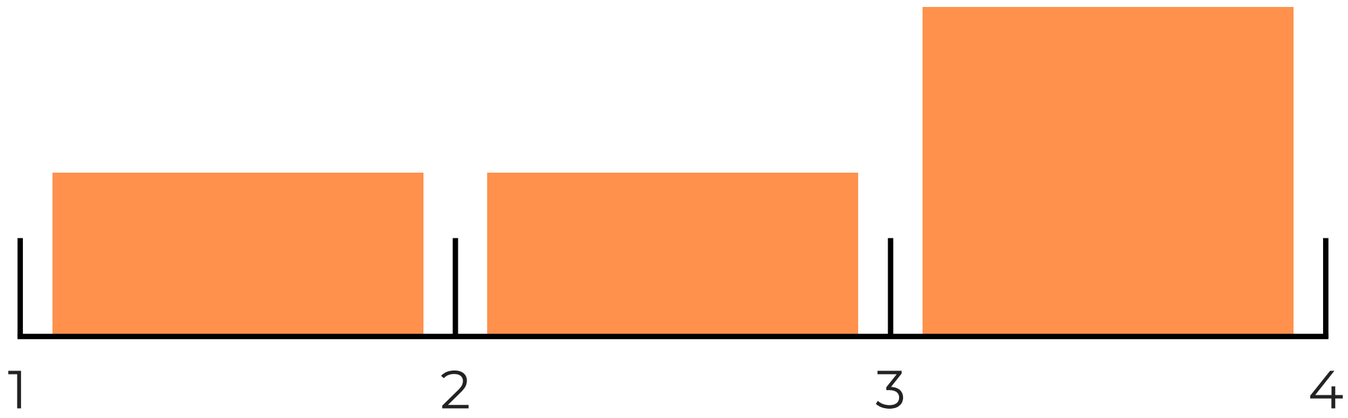
$$bin_i = min\{2, \lfloor 3 \times \frac{4 - 1}{4 - 1} \rfloor\} = bin_2$$



Problem 1: Aggregate Items

Calculating a Continuous Histogram

$$X = \{1, 2.5, 3, 4\}$$

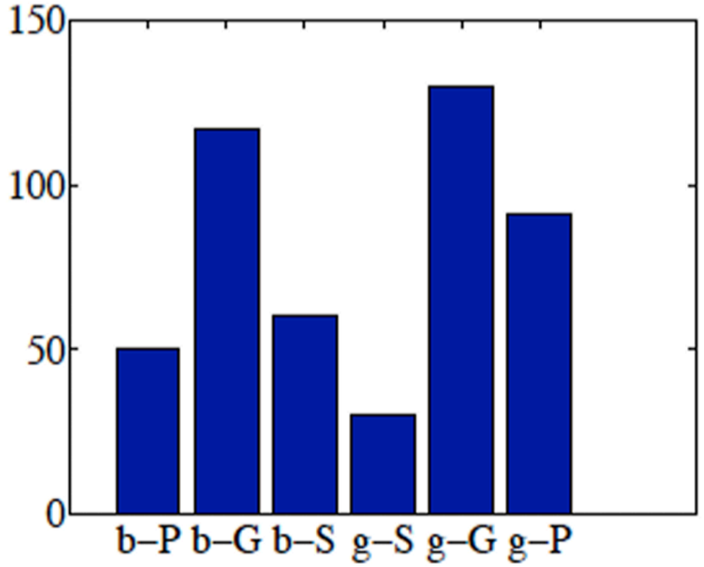


Problem 1: Aggregate Items

Histogram (**2D categorical**):

- Bar Chart

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

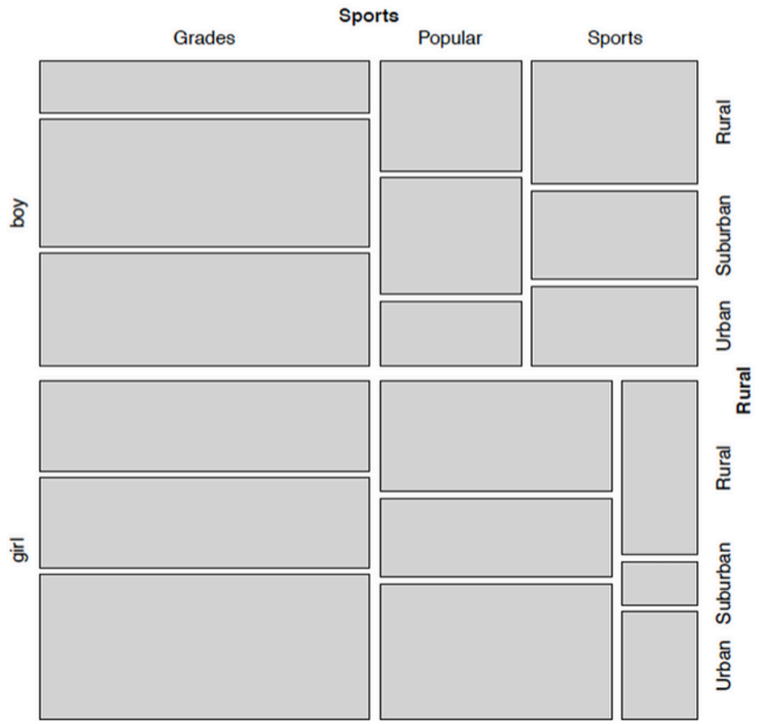


Problem 1: Aggregate Items

Histogram (**2D categorical**):

- Mosaic Plot

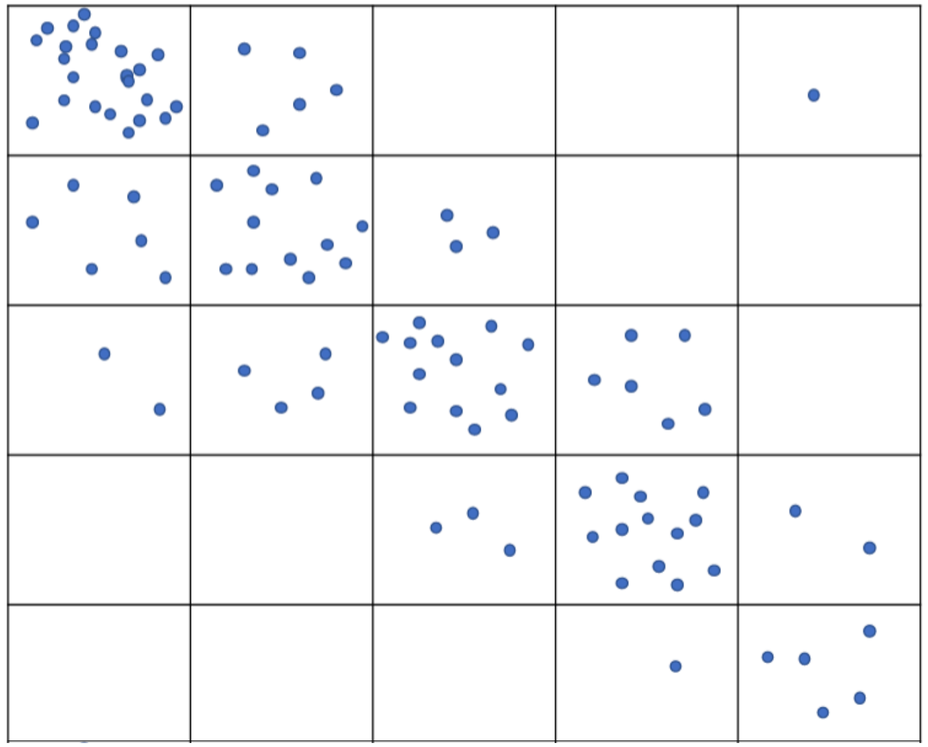
Gender	Goal	Gender	Goal
boy	Sports	girl	Sports
boy	Popular	girl	Grades
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	girl	Grades
girl	Popular	girl	Sports
girl	Grades	girl	Popular
girl	Sports	girl	Grades
girl	Sports	girl	Sports



Problem 1: Aggregate Items

Histogram (**2D ordinal**):

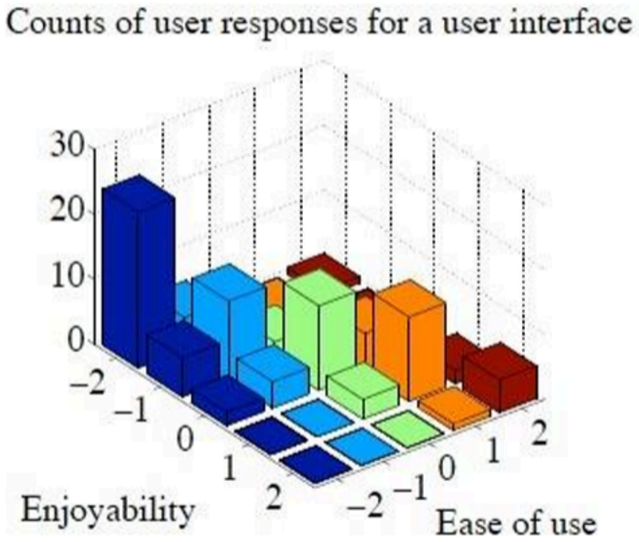
		Ease of use				
		-2	-1	0	1	2
Enjoyability	-2	24	5	0	0	1
	-1	6	12	3	0	0
	0	2	4	13	6	0
	1	0	0	3	13	2
	2	0	0	0	1	5

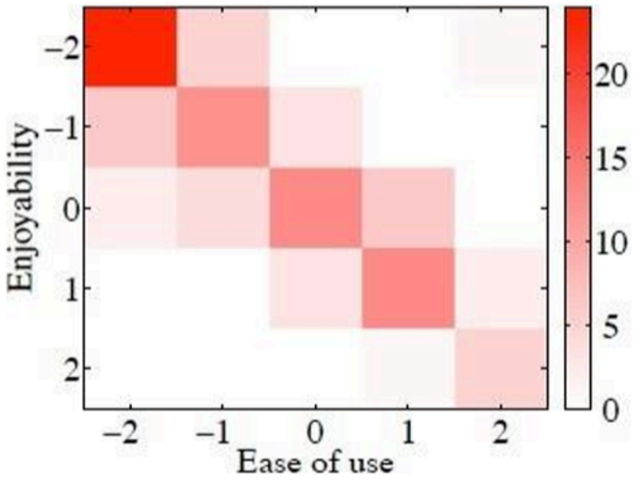


Problem 1: Aggregate Items

Histogram (**2D ordinal**):

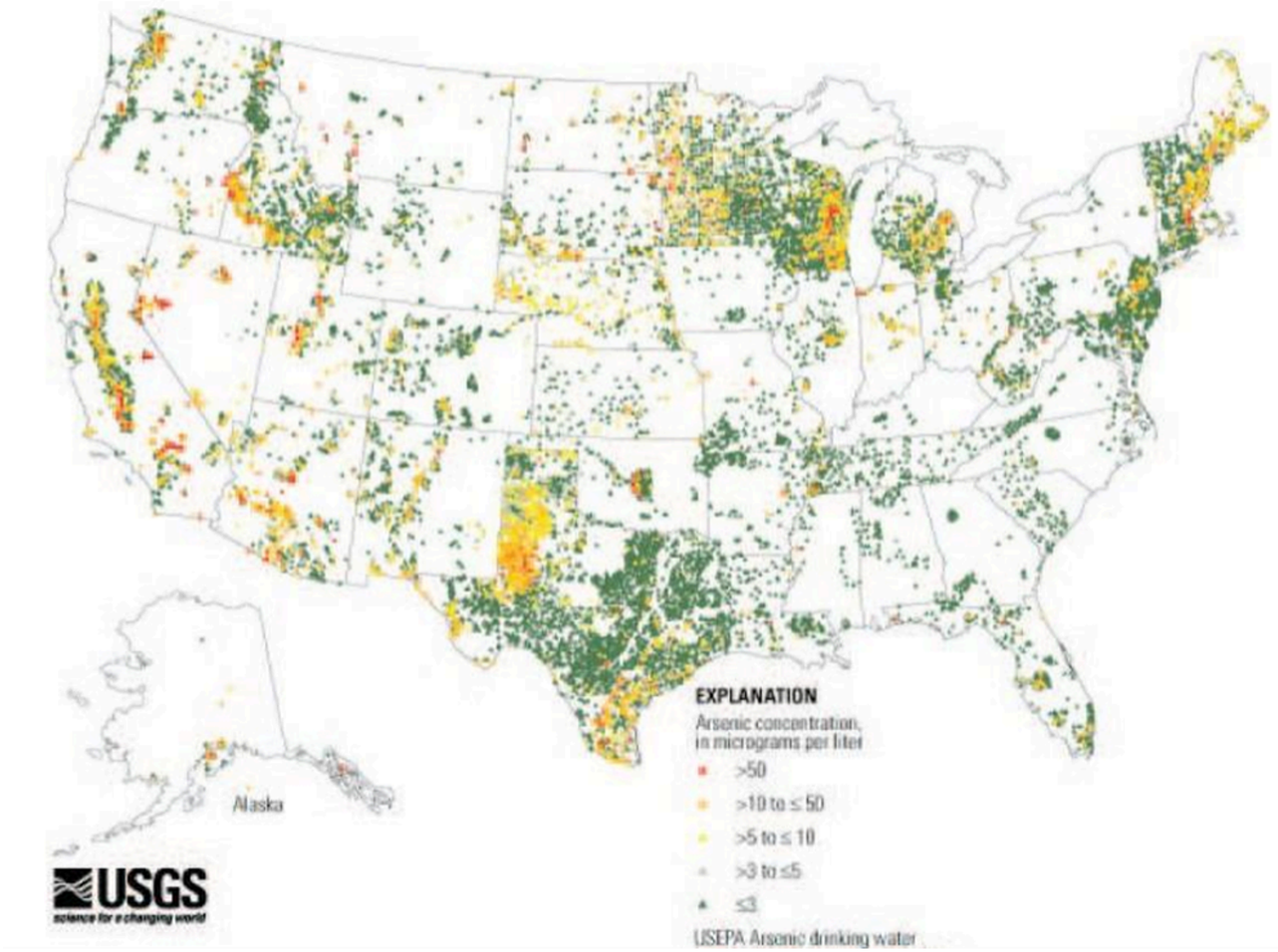
Enjoyability \ Ease of use	-2	-1	0	1	2
-2	24	5	0	0	1
-1	6	12	3	0	0
0	2	4	13	6	0
1	0	0	3	13	2
2	0	0	0	1	5





Problem 1: Aggregate Items

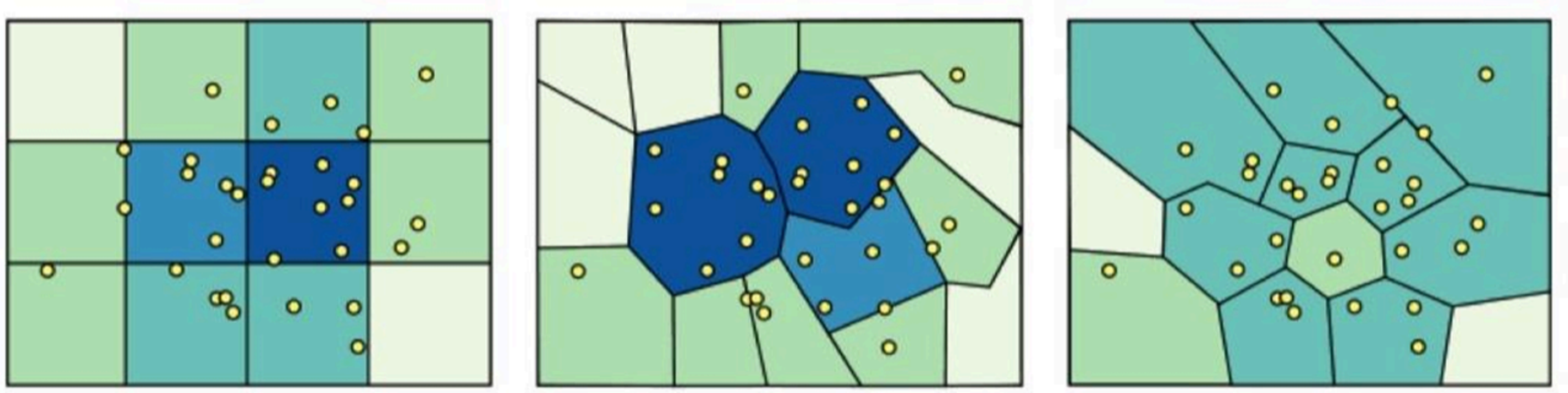
Geographical Data



Problem 1: Aggregate Items

Geographical Data

- Be careful choosing bin boundaries



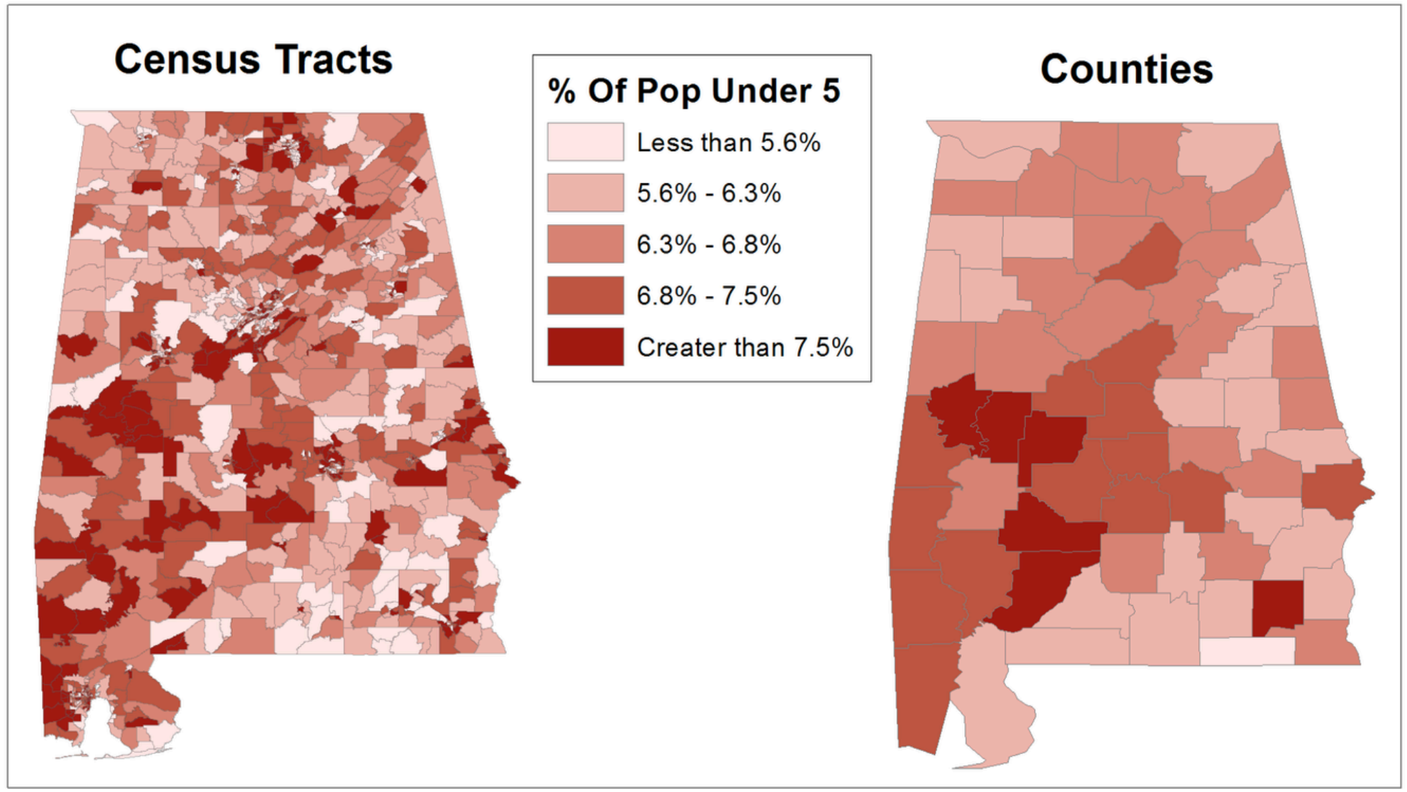
Modifiable Areal Unit Problem:

- which boundaries you choose can dramatically affect results

Problem 1: Aggregate Items

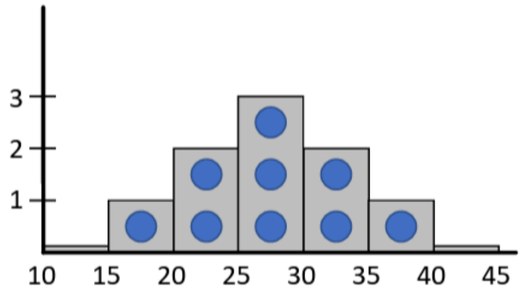
Geographical Data

- Be careful choosing bin boundaries

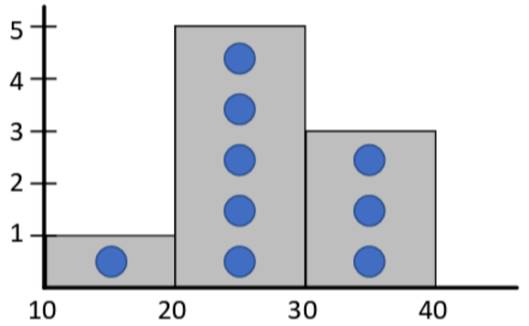


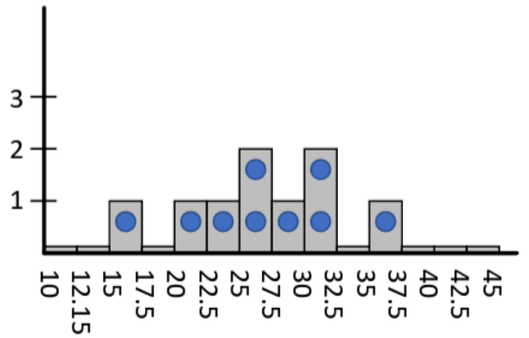
Problem 1: Aggregate Items

The resolution of a histogram can dramatically affect its visual representation



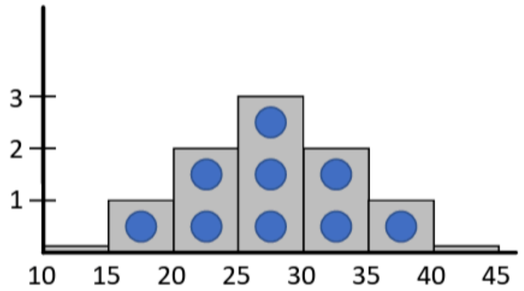
Mean (Average) = 27
Standard Deviation = 6



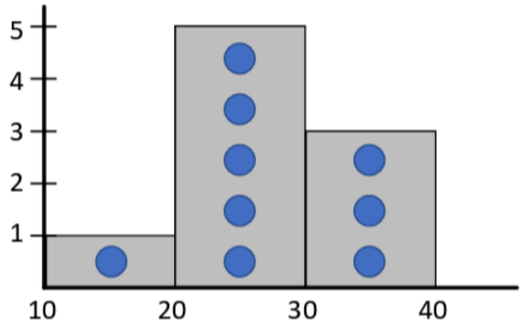


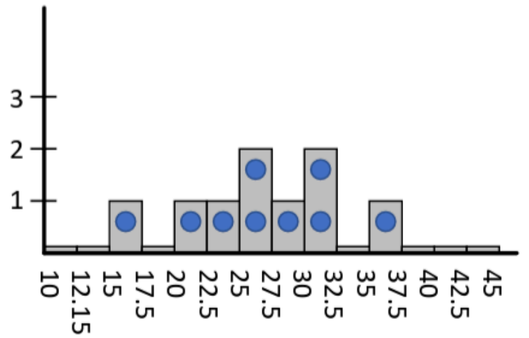
Problem 1: Aggregate Items

The resolution of a histogram can dramatically affect its visual representation



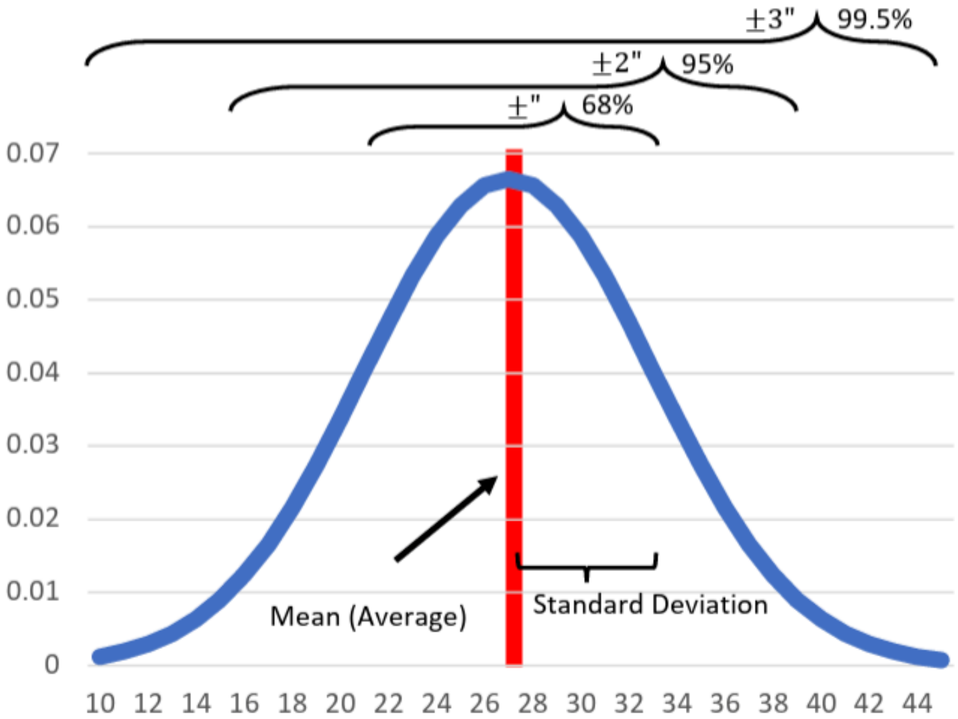
Mean (Average) = 27
Standard Deviation = 6





Problem 1: Aggregate Items

A possible solution: compute statistical measures and visualize the distribution



Problem 1: Aggregate Items

Mean:

Definition: 3.1 *Mean*

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

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

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

Problem 1: Aggregate Items

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
- the average distance of points from the mean

Problem 1: Aggregate Items

Standard Coordinates (or Z-Score):

Definition: 3.8 *Standard coordinates*

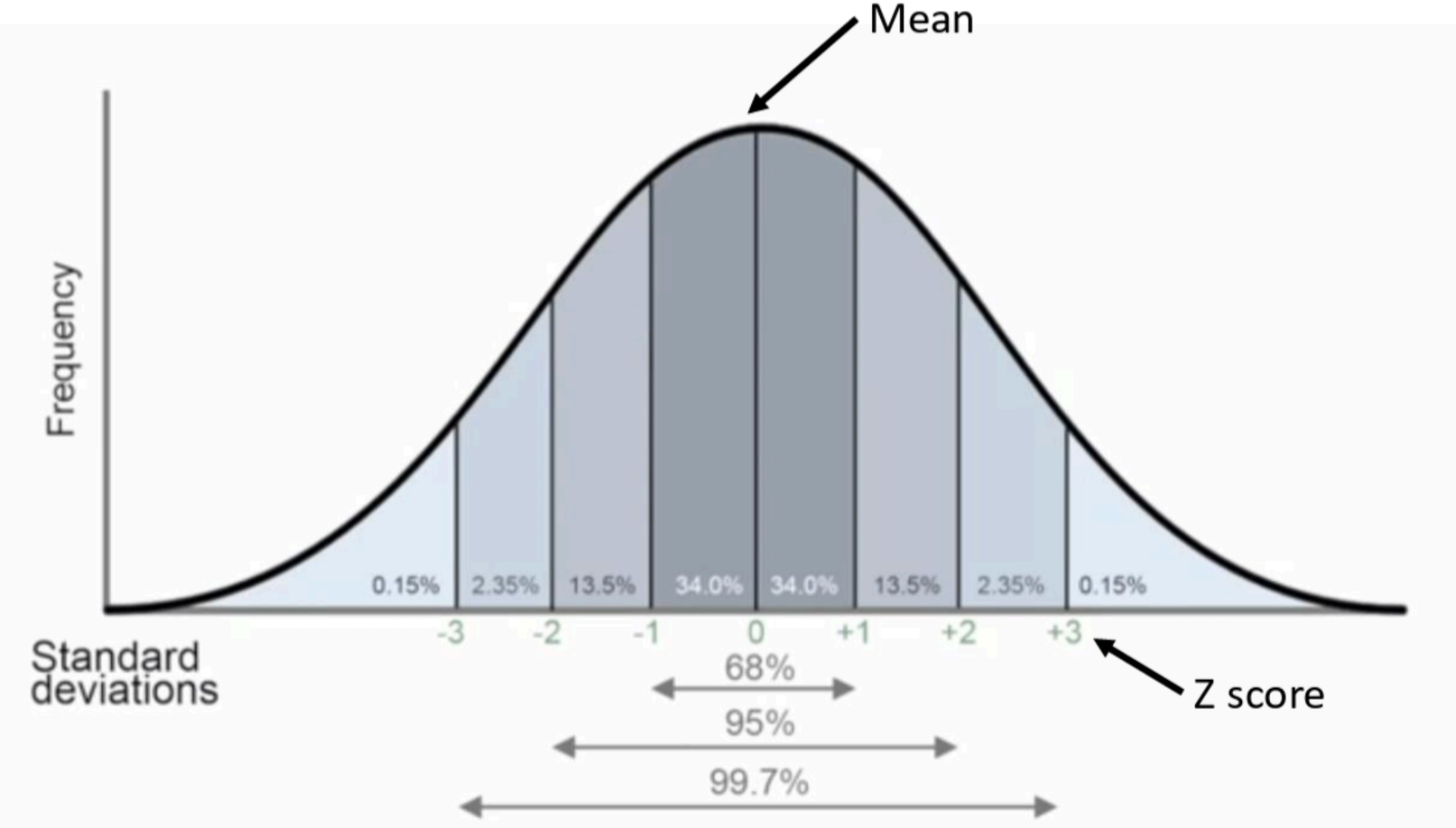
Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. We represent these data items in standard coordinates by computing

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

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

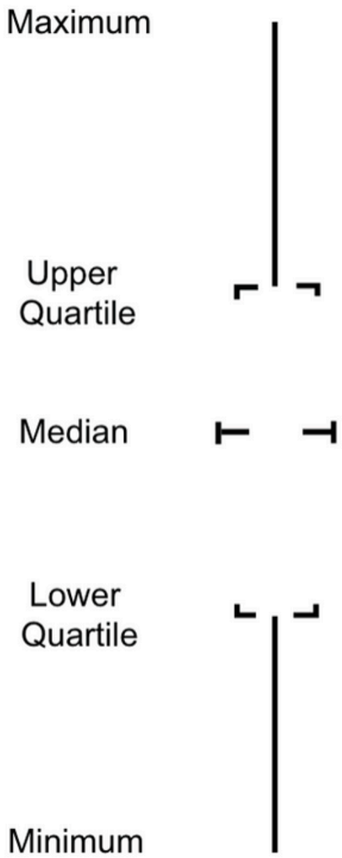
- number of standard deviations away a point is from the mean

Problem 1: Aggregate Items



Problem 1: Aggregate Items

Sometimes, our data aren't normally distributed...

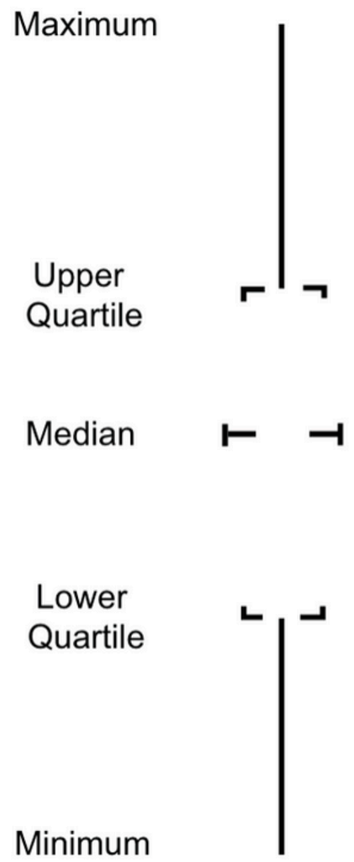


...in these cases, we can use boxplots

Problem 1: Aggregate Items

Boxplots:

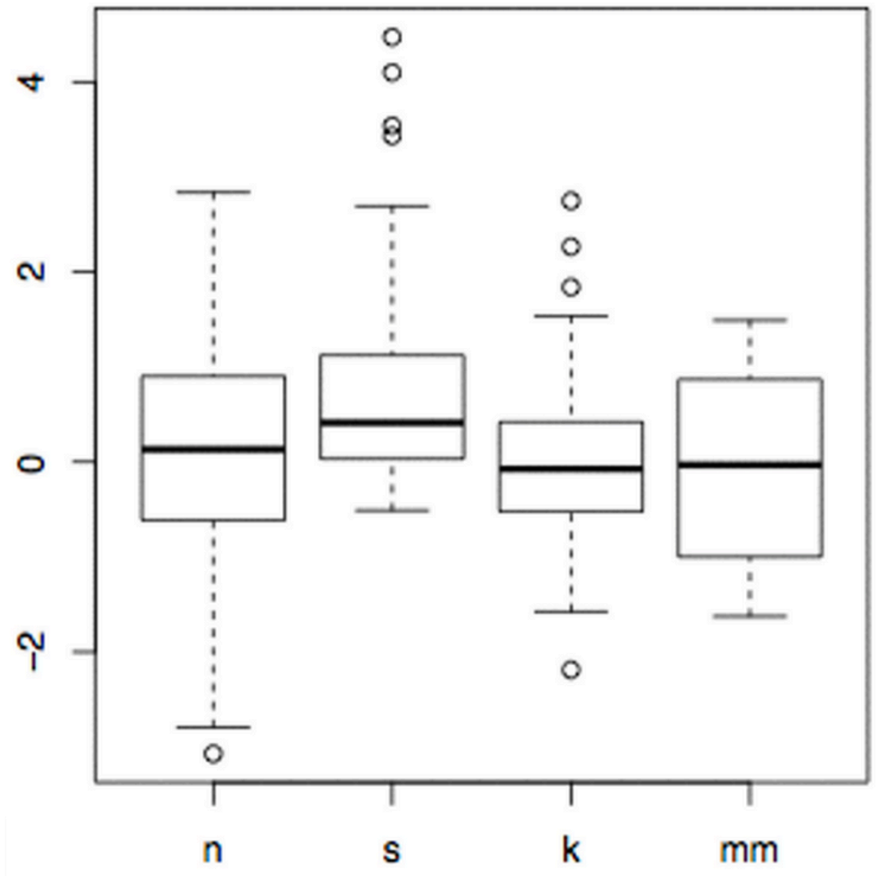
- **median:** center point of your sorted data
- **first quartile (Q1):** center point of the lower half of your sorted data
- **third quartile (Q3):** center point of the upper half of your sorted data



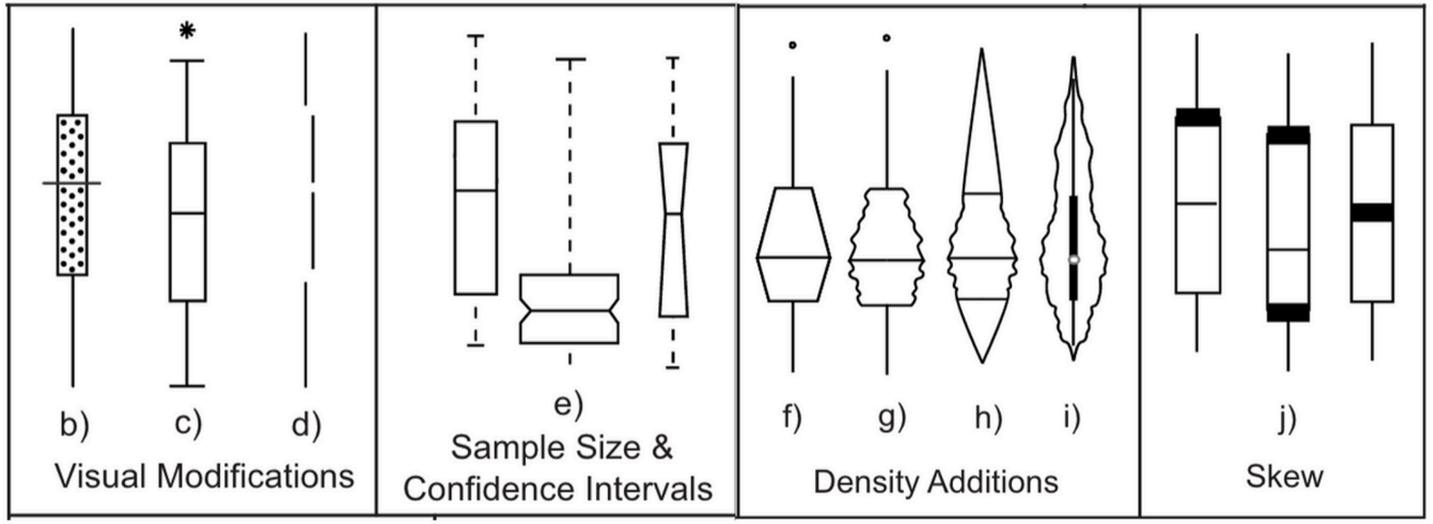
Problem 1: Aggregate Items

Boxplots:

- **Interquartile Range (IQR) = Q3 - Q1**
- **upper bound** is sometimes drawn as the maximum point **<= Q3 + 1.5*IQR**
- **lower bound** is sometimes drawn as the minimum point **>= Q1 - 1.5*IQR**



Problem 1: Aggregate Items



Eurographics/ IEEE-VGTC Symposium on Visualization 2010
G. Melançon, T. Munzner, and D. Weiskopf
(Guest Editors)

Volume 29 (2010), Number 3

Visualizing Summary Statistics and Uncertainty

K. Potter[1], J. Kniss[2], R. Riesenfeld[3], and C.R. Johnson[1]

[1]Scientific Computing and Imaging Institute, University of Utah
[2]Department of Computer Science, University of New Mexico
[3]School of Computing, University of Utah

Abstract

The graphical depiction of uncertainty information is emerging as a problem of great importance. Scientific data sets are not considered complete without indications of error, accuracy, or levels of confidence. The visual portrayal of this information is a challenging task. This work takes inspiration from graphical data analysis to create visual representations that show not only the data value, but also important characteristics of the data including uncertainty. The canonical box plot is reexamined and a new hybrid summary plot is presented that incorporates a collection of descriptive statistics to highlight salient features of the data. Additionally, we present an extension of the summary plot to two dimensional distributions. Finally, a use-case of these new plots is presented, demonstrating their ability to present high-level overviews as well as detailed insight into the salient features of the underlying data distribution.

Categories and Subject Descriptors (according to ACM CCS): I.3.6 [Computer Graphics]: Methodology and Techniques

1. Introduction

As computational power, memory limits, and bandwidth have inexorably increased, so has the corresponding size and complexity of the data sets generated by scientists. Because of the reduction of hardware limitations, simulations can be run at higher resolutions, for longer amounts of time, using more sophisticated numerical models. We can compute more exhaustively, store more abundantly, and access data more rapidly, all of which leads researchers to create more complex data sets. Visualization approaches overlook available uncertainty information [JS03, Joh04]. As the importance of visualizing these large, complex data sets grows, the actual task of visualizing them becomes more complicated; incorporating the additional data parameter of uncertainty into the visualizations becomes even less straightforward. Difficulties in applying preexisting methods, additional visual clutter, and the lack of obvious visualization techniques leave uncertainty visualization an unsolved problem.

https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0b26efd3ff59bf58f1f2dc41a190147b6ec0846b

Problem 2: Aggregate Attributes

Two goals:

- Group attributes and compute similarity across the set of items
- Conduct dimensionality reduction to preserve meaningful structure

Problem 2: Aggregate Attributes

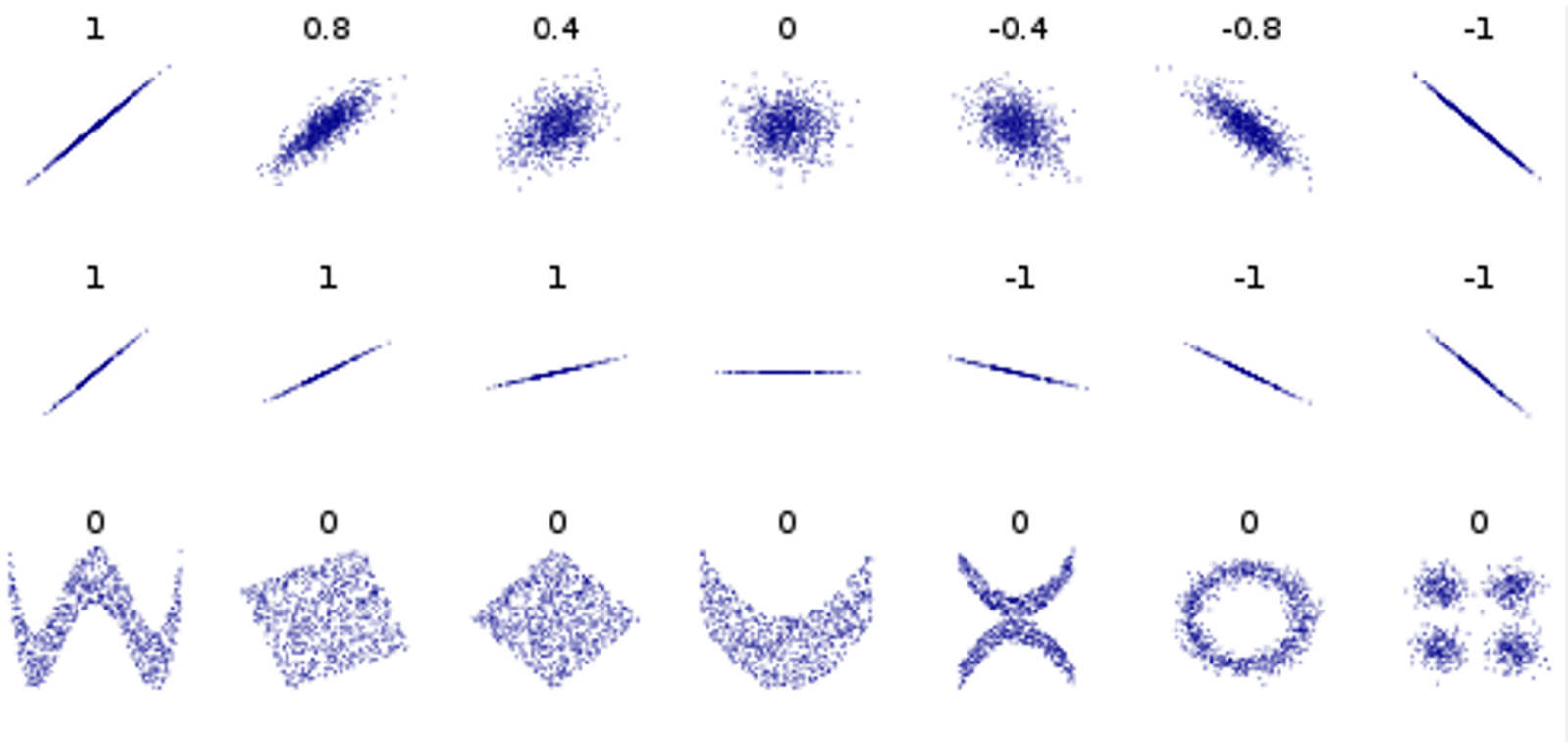
Similarity Scores:

- Correlation:
  - measure of similarity between 2 or more attributes
  - many variations (e.g., pearson, spearman, etc.)
- Regression:
  - fit a model to the data
  - measure the quality of fit

Problem 2: Aggregate Attributes

Pearson Correlation Coefficient:

- A measure of linearity between 2 sets



Problem 2: Aggregate Attributes

Pearson Correlation Coefficient:

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

Problem 2: Aggregate Attributes

Pearson Correlation Coefficient:

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Problem 2: Aggregate Attributes

Pearson Correlation Coefficient:

$$X = \{1, 2.5, 3, 4.5\}$$
$$Y = \{2, 2.5, 3.5, 4\}$$

$$\bar{x} = 2.75$$
$$\bar{y} = 3$$

$$\sigma_X = \sqrt{(1 - 2.75)^2 + (2.5 - 2.75)^2 + (3 - 2.75)^2 + (4.5 - 2.75)^2} = 1.25$$

$$\sigma_Y = \sqrt{(2 - 3)^2 + (2.5 - 3)^2 + (3.5 - 3)^2 + (4 - 3)^2} = 0.79$$

Problem 2: Aggregate Attributes

Pearson Correlation Coefficient:

$$X = \{1, 2.5, 3, 4.5\}$$
$$Y = \{2, 2.5, 3.5, 4\}$$

$$\bar{x} = 2.75$$
$$\bar{y} = 3$$
$$\sigma_X = 1.25$$
$$\sigma_Y = 0.79$$

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_X\sigma_Y}$$

$$= \frac{(1 - 2.75)(2 - 3) + (2.5 - 2.75)(2.5 - 3) + (3 - 2.75)(3.5 - 3) + (4.5 - 2.75)(4 - 3)}{4(1.25)(0.79)}$$

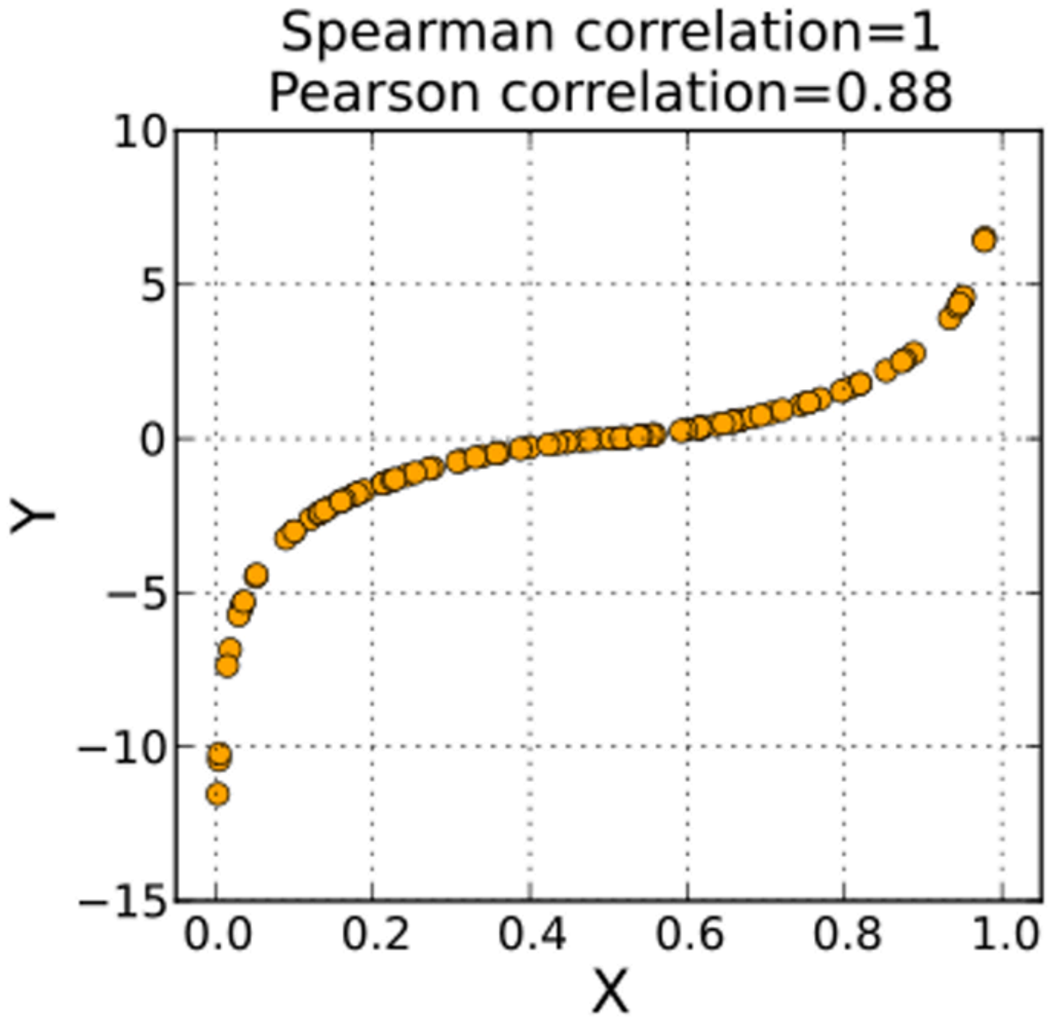
$$= 0.95$$

Problem 2: Aggregate Attributes

What if your data are non-linearly correlated?

Problem 2: Aggregate Attributes

Spearman Rank Correlation



Problem 2: Aggregate Attributes

Spearman Rank Correlation:

IQ, (X)	Hours of TV per week, (Y)
86	0
97	20
99	28
100	27
101	50
103	29
106	7
110	17
112	6
113	12

Problem 2: Aggregate Attributes

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

Problem 2: Aggregate Attributes

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

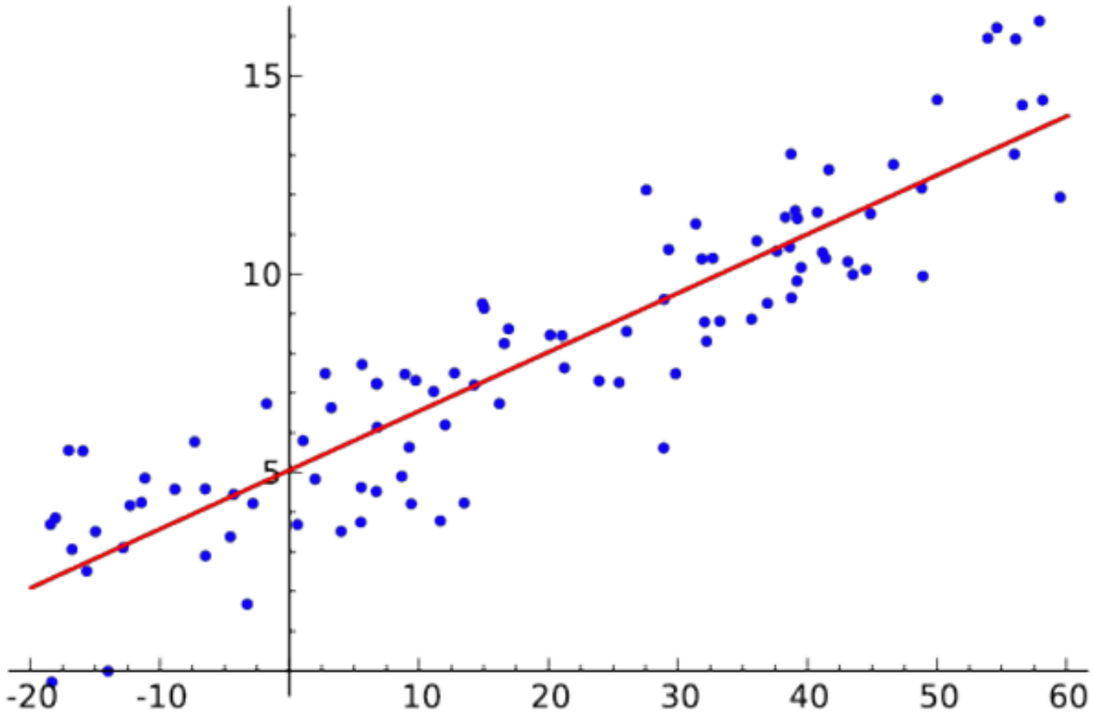
PCC(X',Y') = -0.1758

Problem 2: Aggregate Attributes

Regression: Fitting a model to your data

Given $y_i = \alpha + \beta x_i + \epsilon_i$

Find α and β that minimize ϵ_i



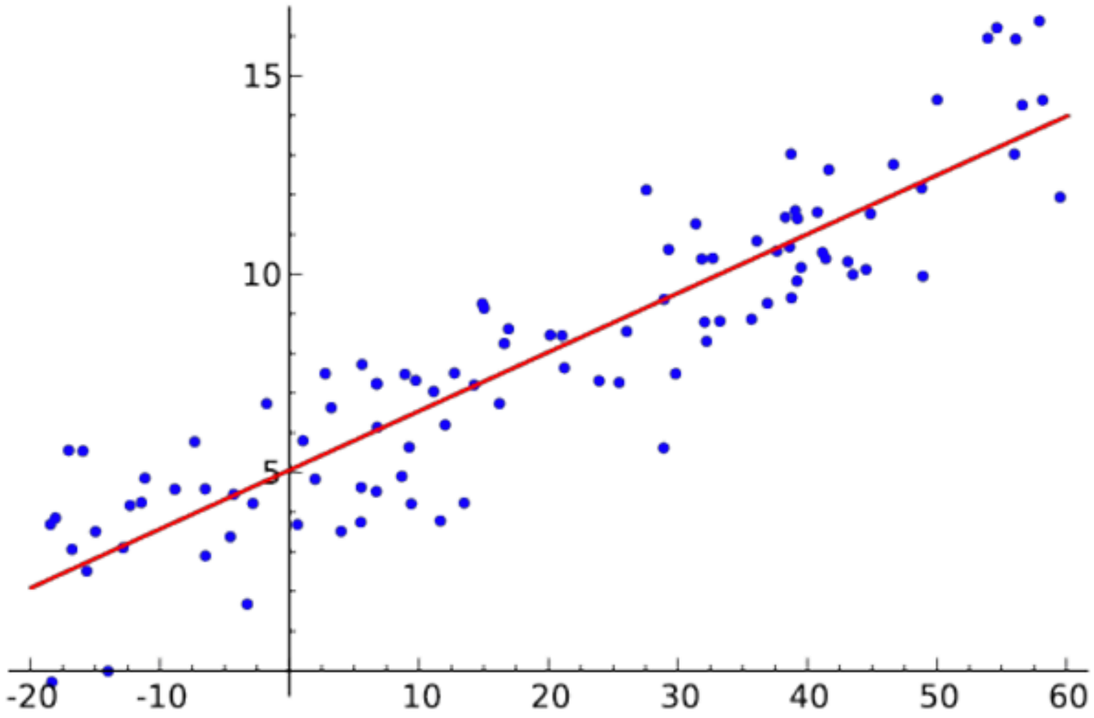
Problem 2: Aggregate Attributes

Regression: Fitting a model to your data

- can be computed directly

$$\hat{\beta} = \frac{cov(X, Y)}{\sigma_X}$$

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



Problem 2: Aggregate Attributes

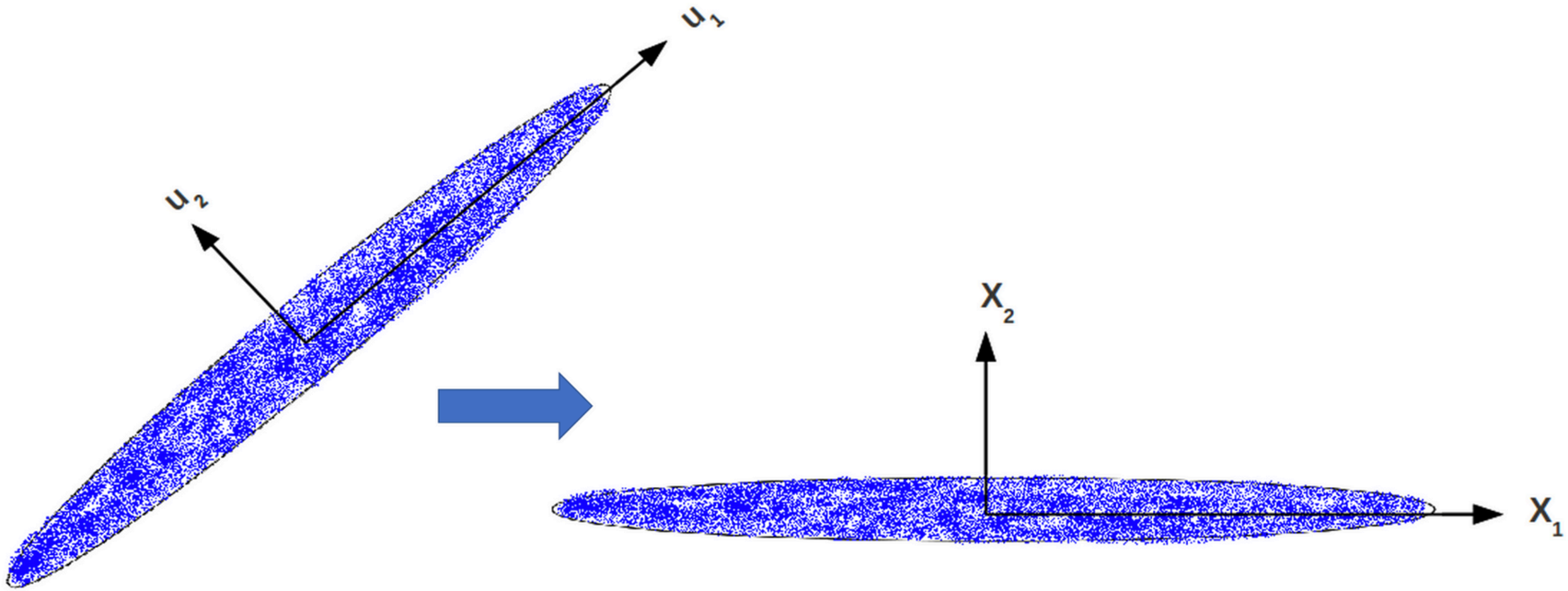
Dimensionality Reduction:

- Reduce the number of attributes in your dataset to be more manageable

Problem 2: Aggregate Attributes

Dimensionality Reduction (Linear):

- Principal Component Analysis (PCA):
  - create new attributes that are linear combinations of the attributes in your dataset



Problem 2: Aggregate Attributes

Dimensionality Reduction (Nonlinear):

- Multidimensional Scaling (MDS):
  - aims to preserve the pairwise distance between items when projecting to lower dimensional space

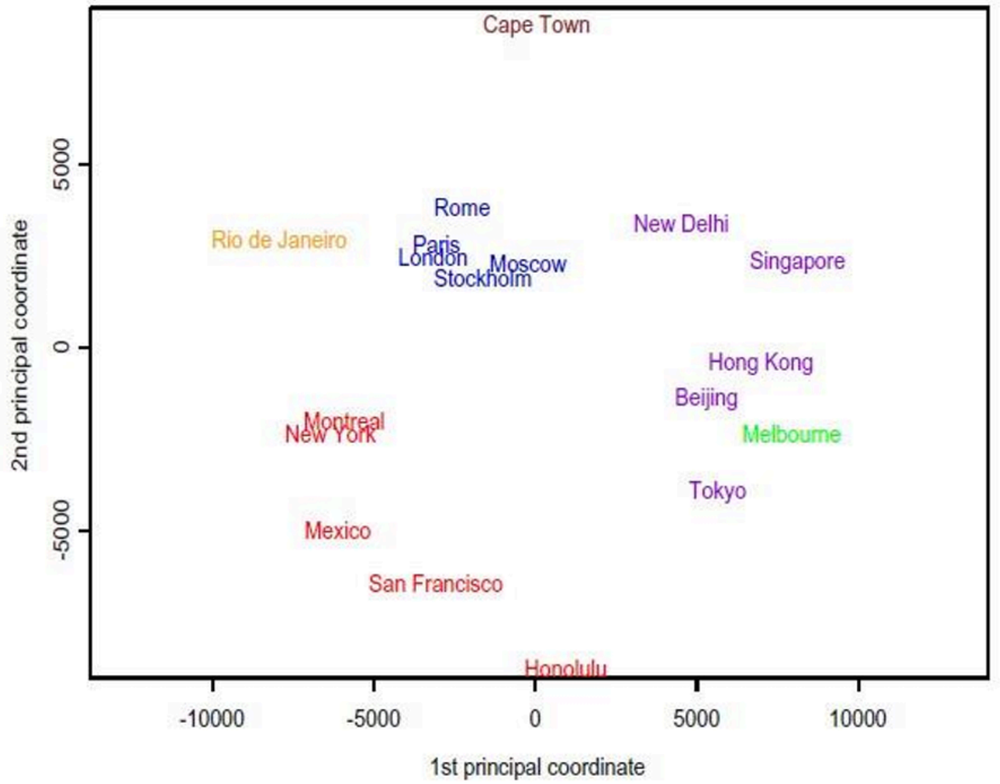
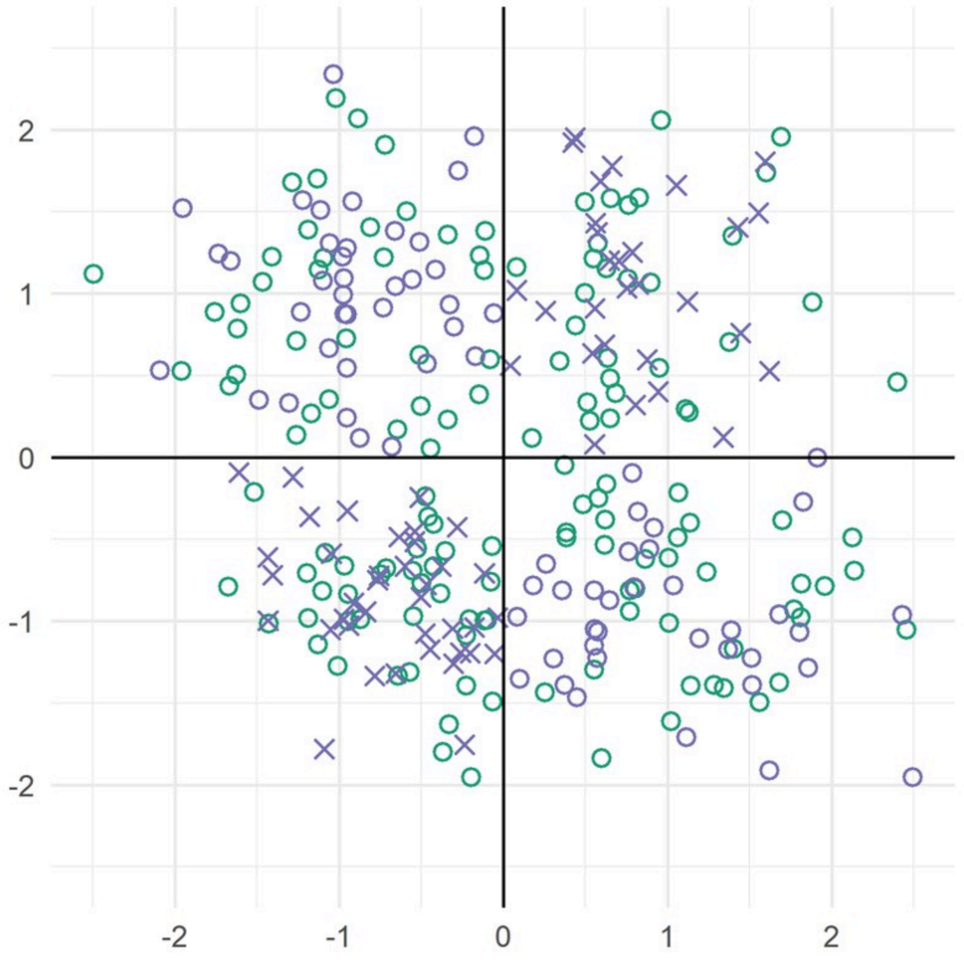


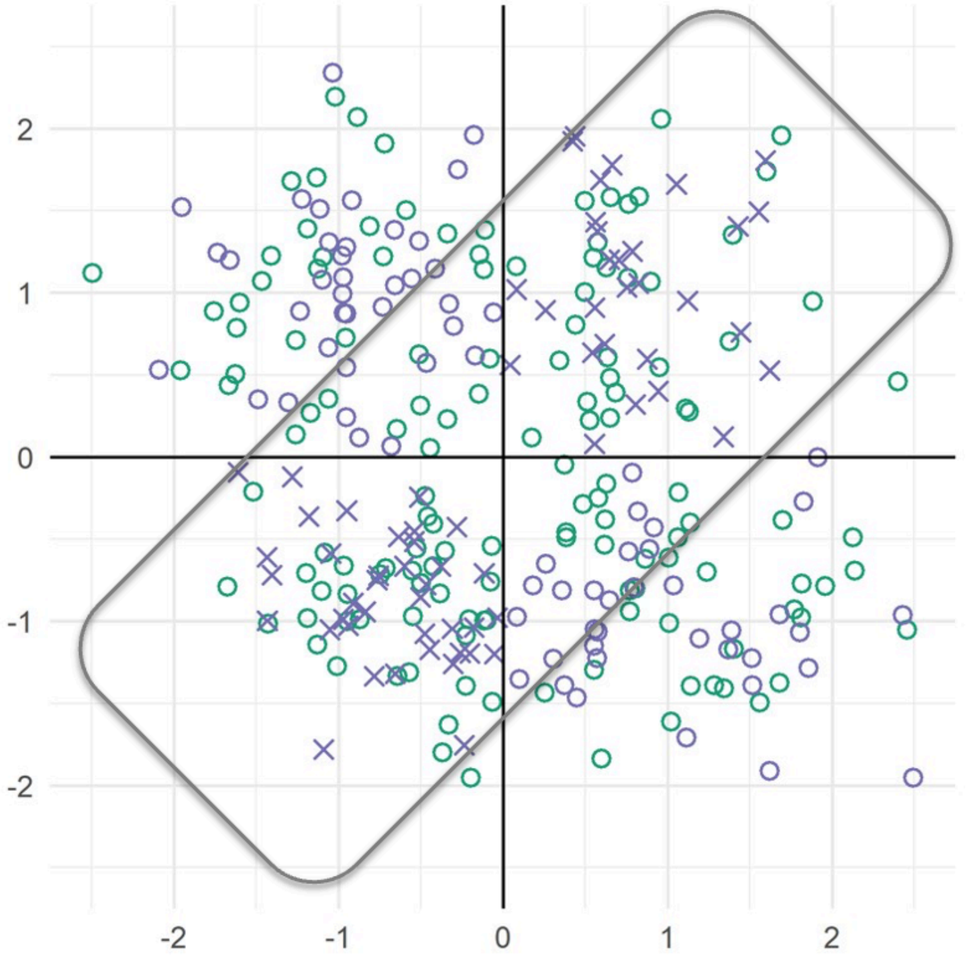
FIGURE 13.1. *Two-dimensional map of 18 world cities using the classical scaling algorithm on airline distances between those cities. The colors*

Problem 3: What is Lost or Misinterpreted?

I will show you a scatterplot, and you will pay attention to the **blue X's.**

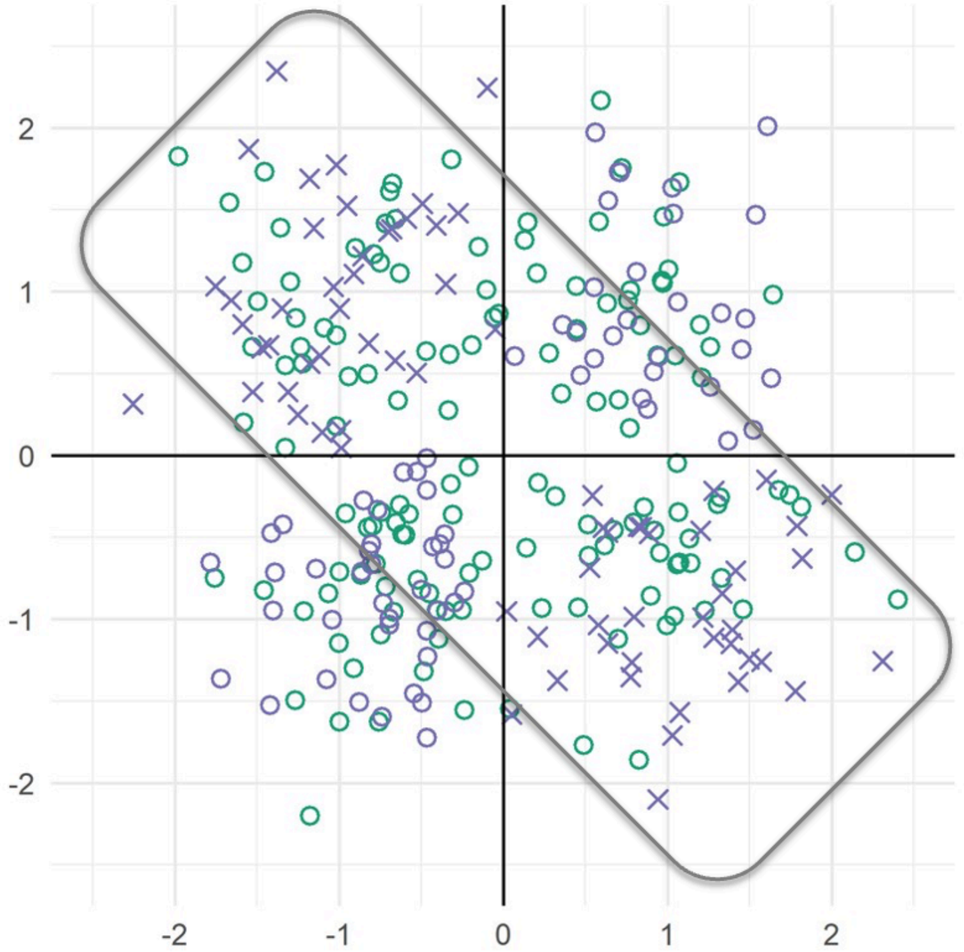


In this sequence, are there more plots where the **blue X's** form a **forward-slash (/)** or a **backward-slash (\)**?



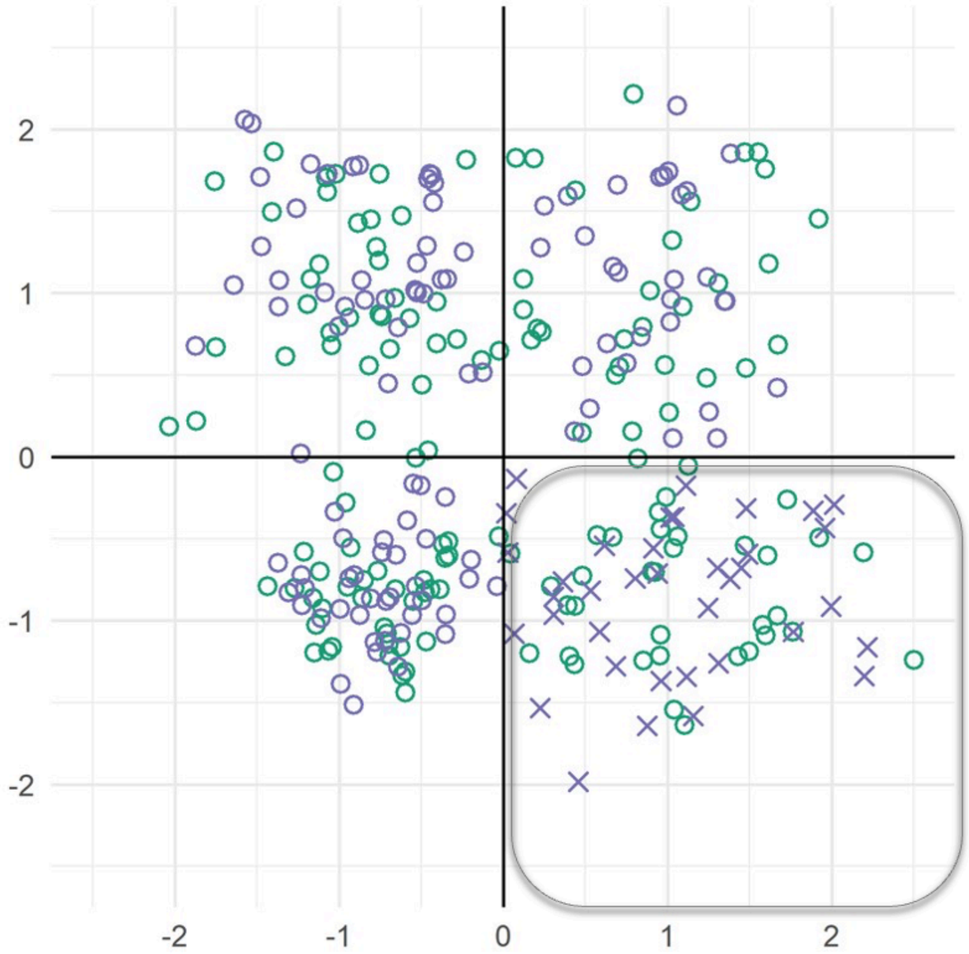
This is a **forward-slash (/)** plot

In this sequence, are there more plots where the **blue X's** form a **forward-slash (/)** or a **backward-slash (\)**?



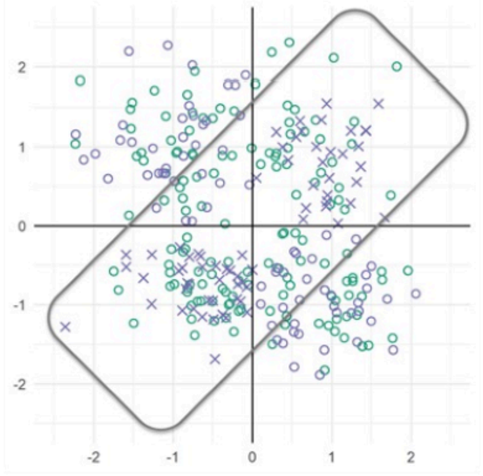
This is a **backward-slash (\)** plot

In this sequence, are there more plots where the **blue X's** form a **forward-slash (/)** or a **backward-slash (\)**?



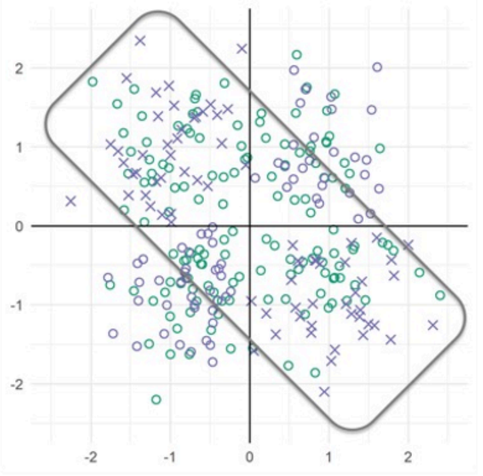
This is **neither (ignore!)**

Count the number of forward and backward slashes.



This is a **forward-slash (/)** plot.

?

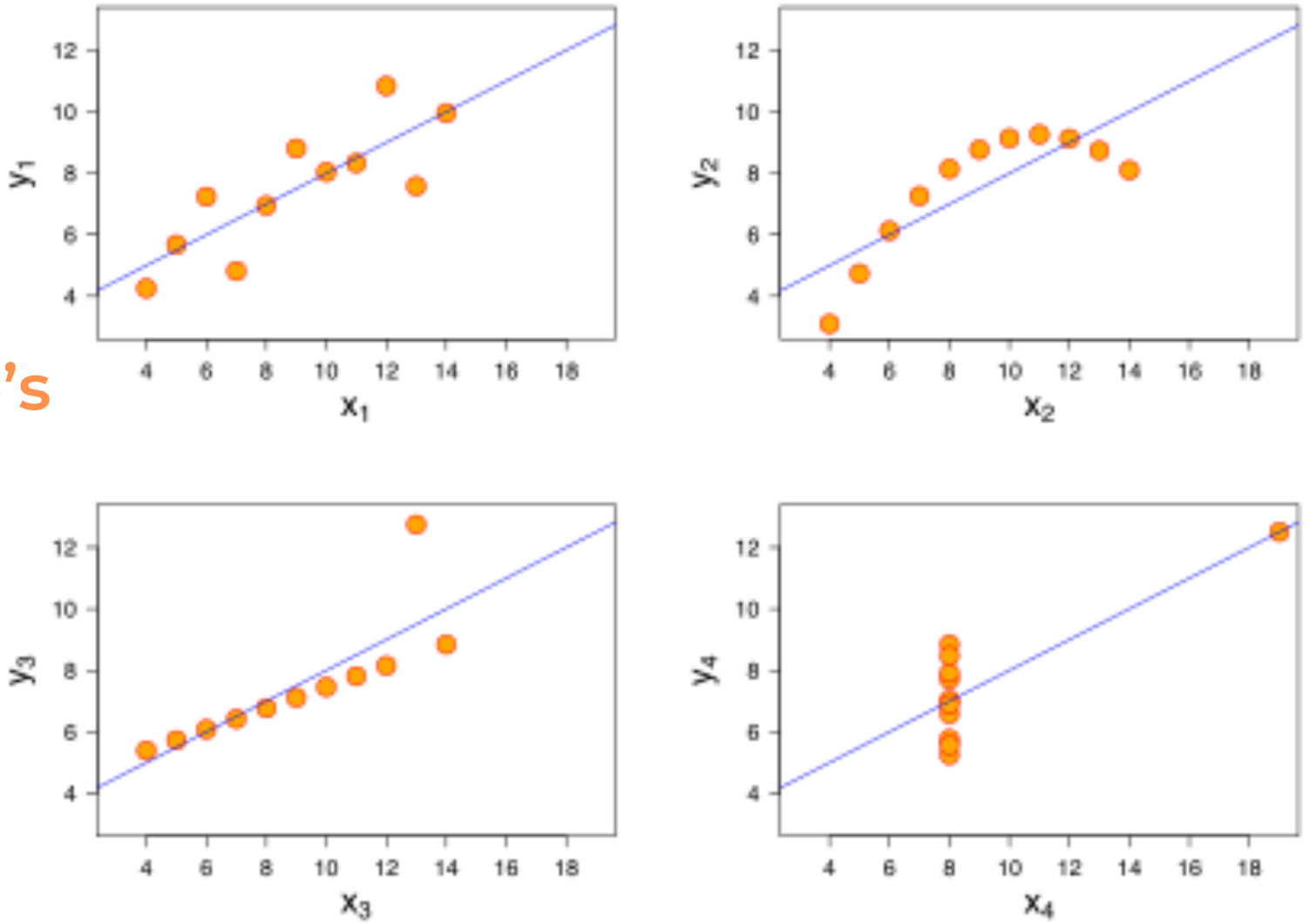


This is a **backward-slash (\)** plot

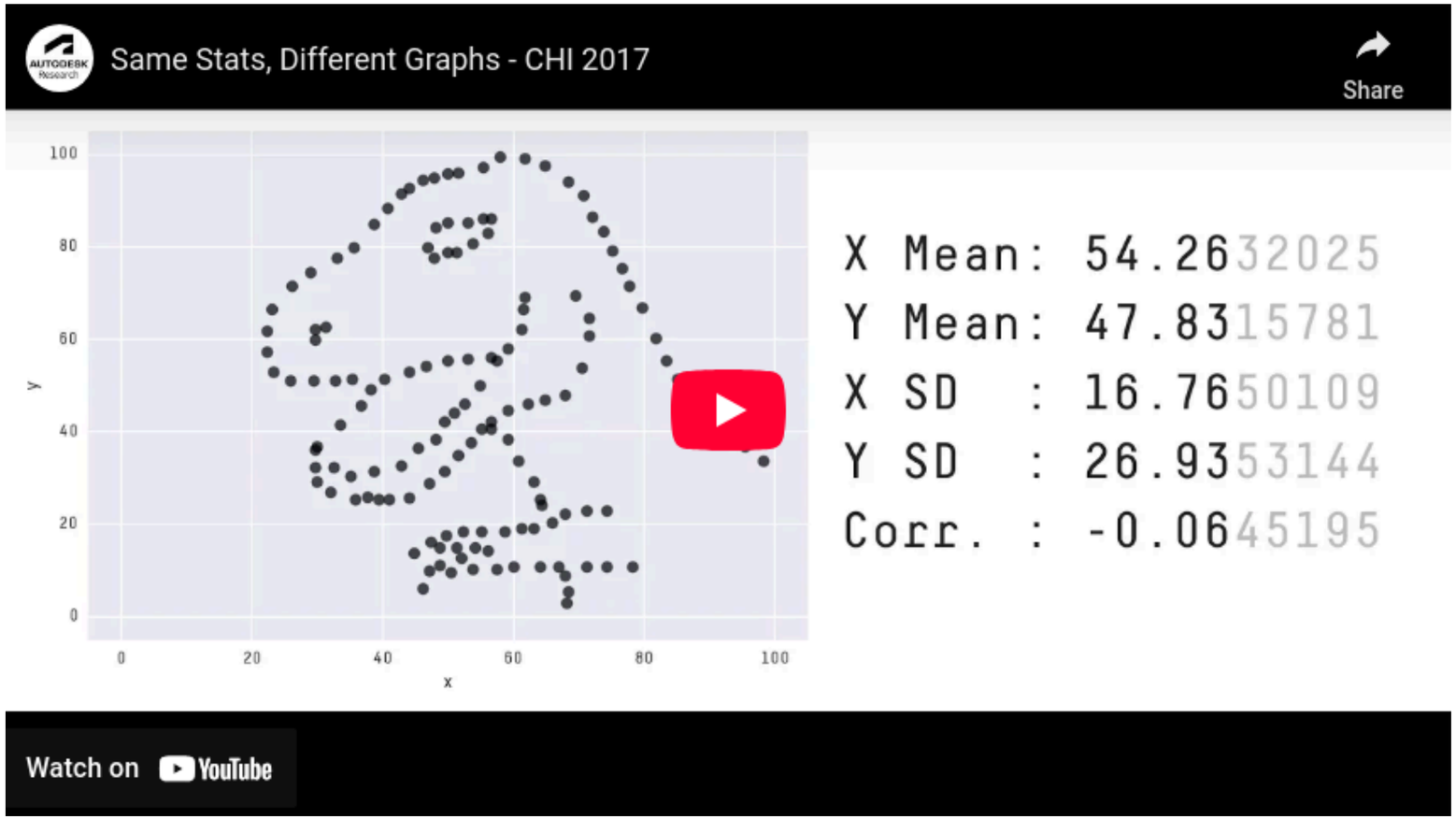
?

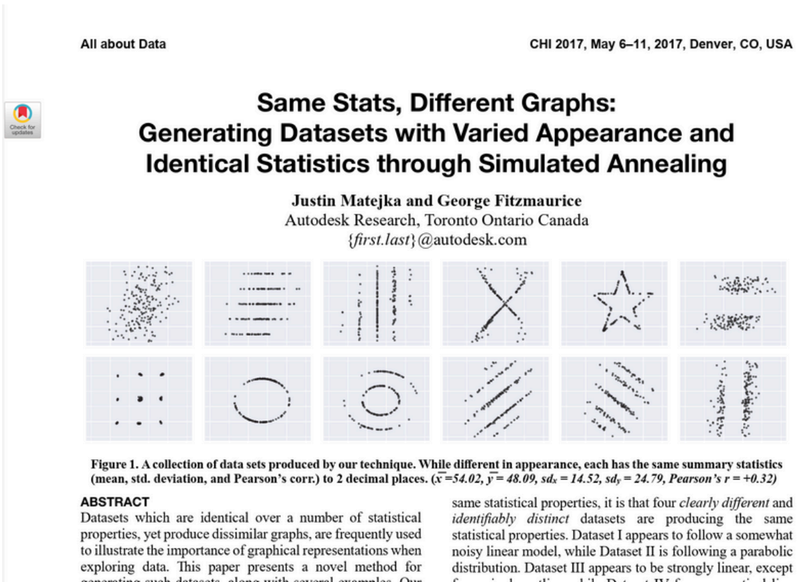
Problem 3: What is Lost or Misinterpreted?

Anscombe's Quartet



Problem 3: What is Lost or Misinterpreted?





All about Data — CHI 2017, May 6–11, 2017, Denver, CO, USA

Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

Justin Matejka and George Fitzmaurice
Autodesk Research, Toronto Ontario Canada
{*first.last*}@autodesk.com

Figure 1. A collection of data sets produced by our technique. While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson's corr.) to 2 decimal places. ($\bar{x}$ =54.02, $\bar{y}$ = 48.09, sd_x = 14.52, sd_y = 24.79, *Pearson's r* = +0.32)

ABSTRACT
Datasets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This paper presents a novel method for

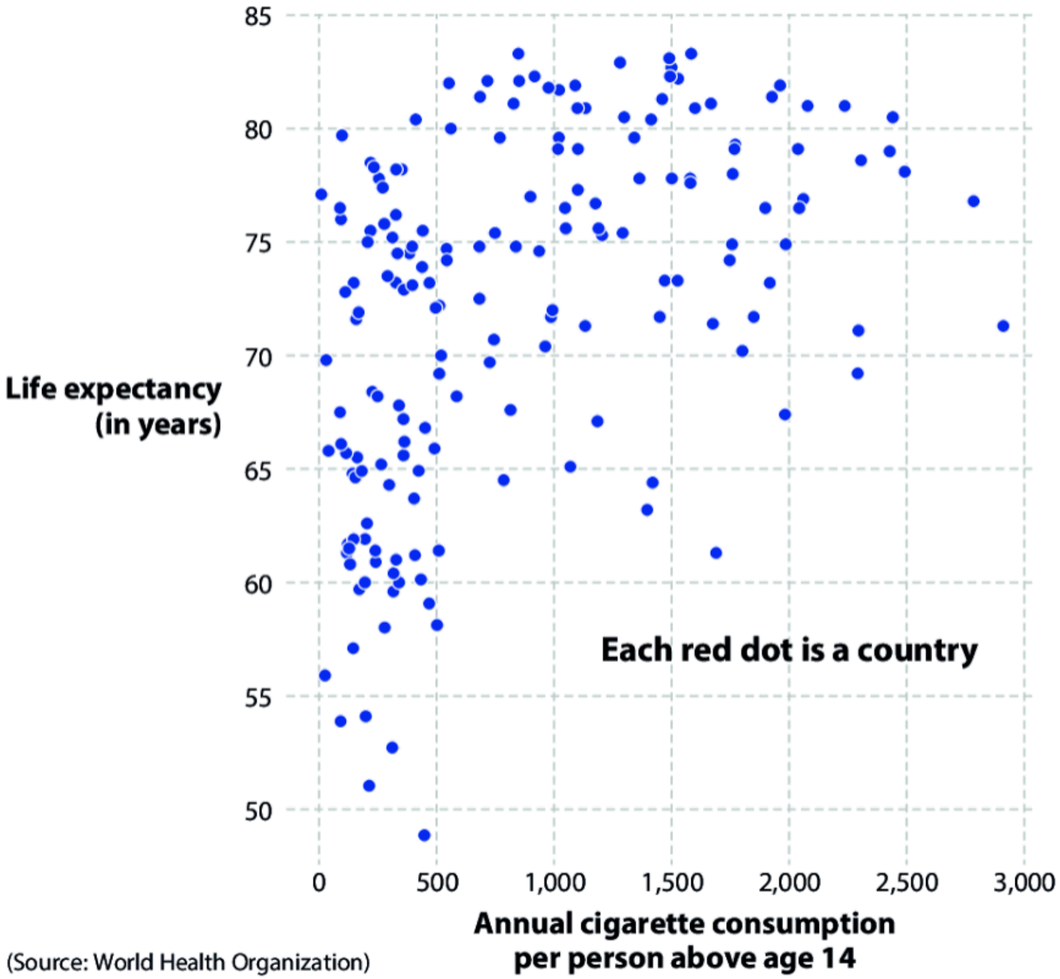
same statistical properties. Dataset I appears to follow a somewhat noisy linear model, while Dataset II is following a parabolic distribution. Dataset III appears to be strongly linear, except

https://dl.acm.org/doi/10.1145/3025453.3025912

Problem 3: What is Lost or Misinterpreted?

“The more cigarettes we smoke, the longer we live!”

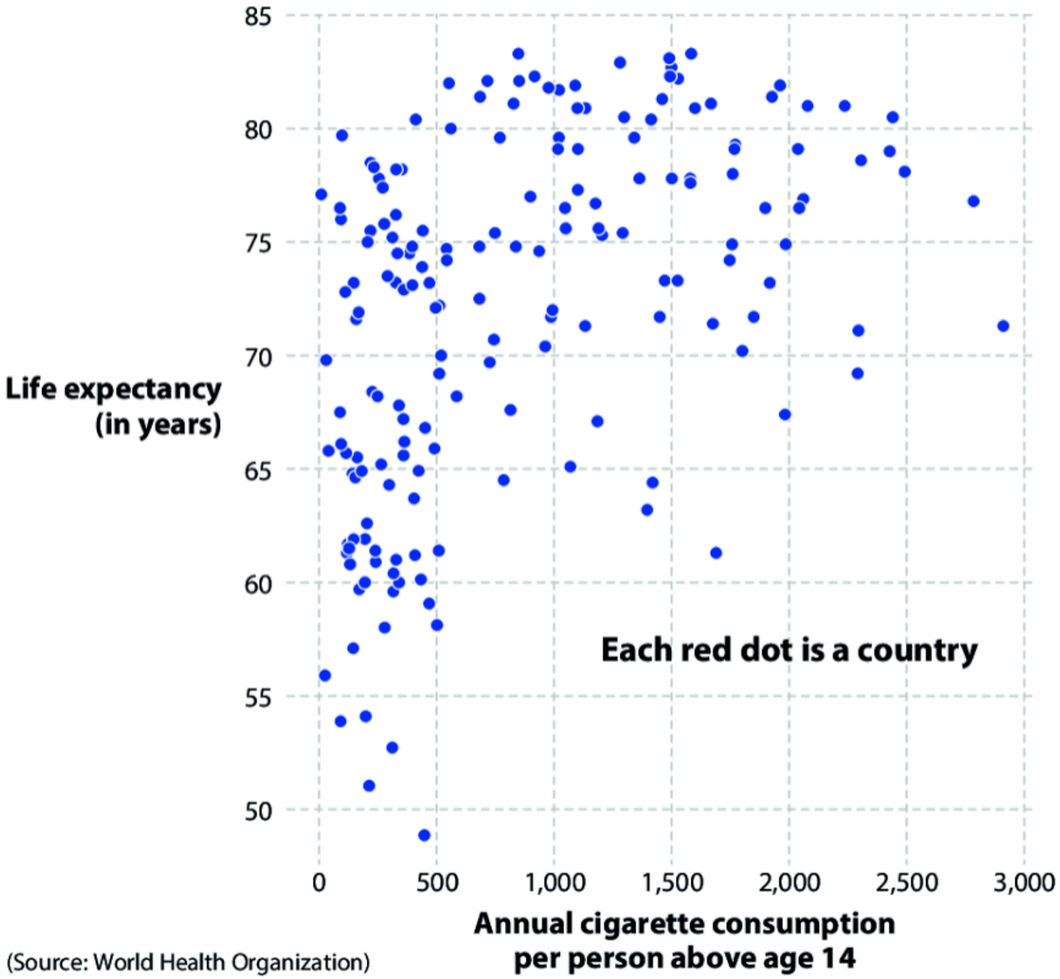
“There is a positive relationship between cigarette consumption and life expectancy at the country level”



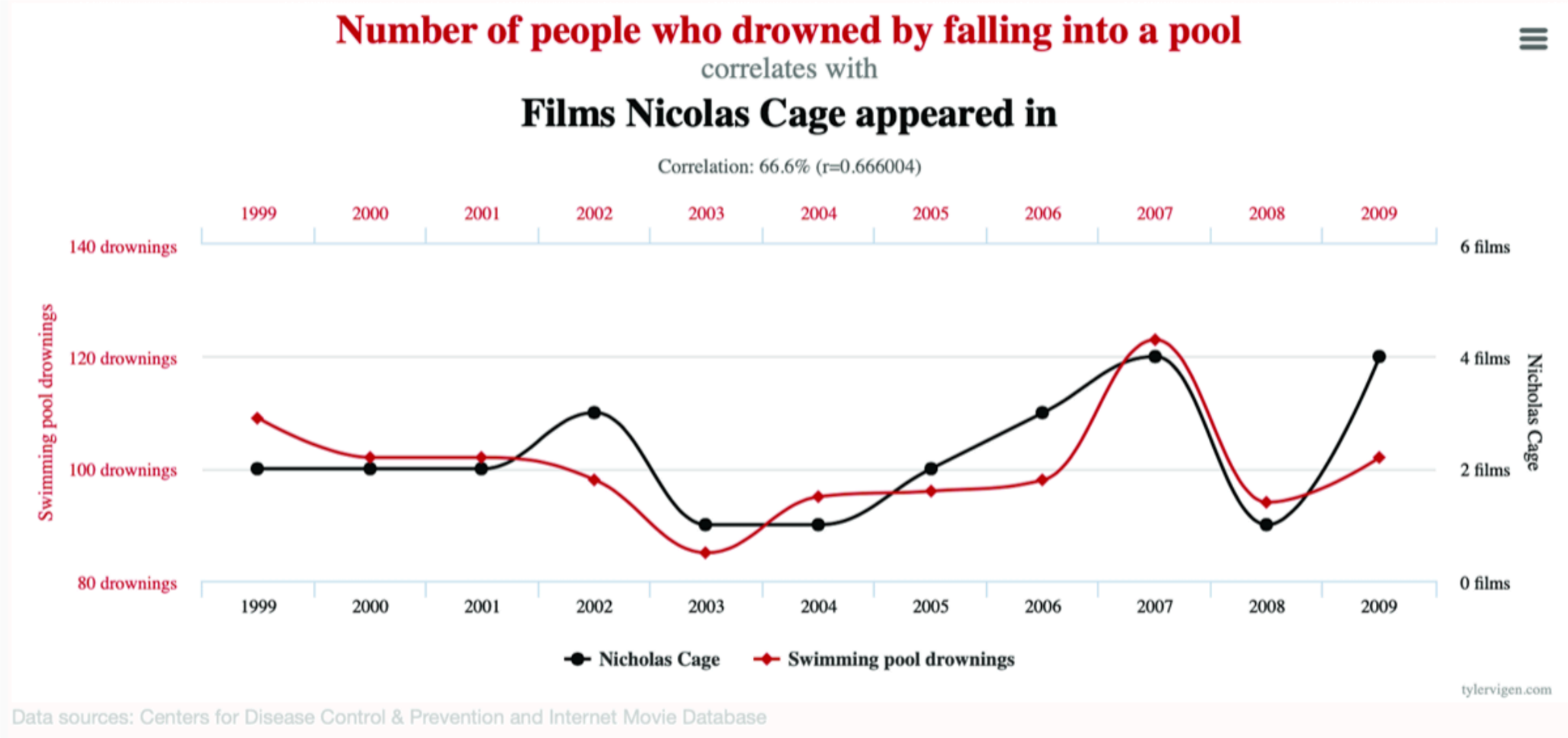
Problem 3: What is Lost or Misinterpreted?

~~"The more cigarettes we smoke, the longer we live!"~~

"There is a positive relationship between cigarette consumption and life expectancy at the country level"

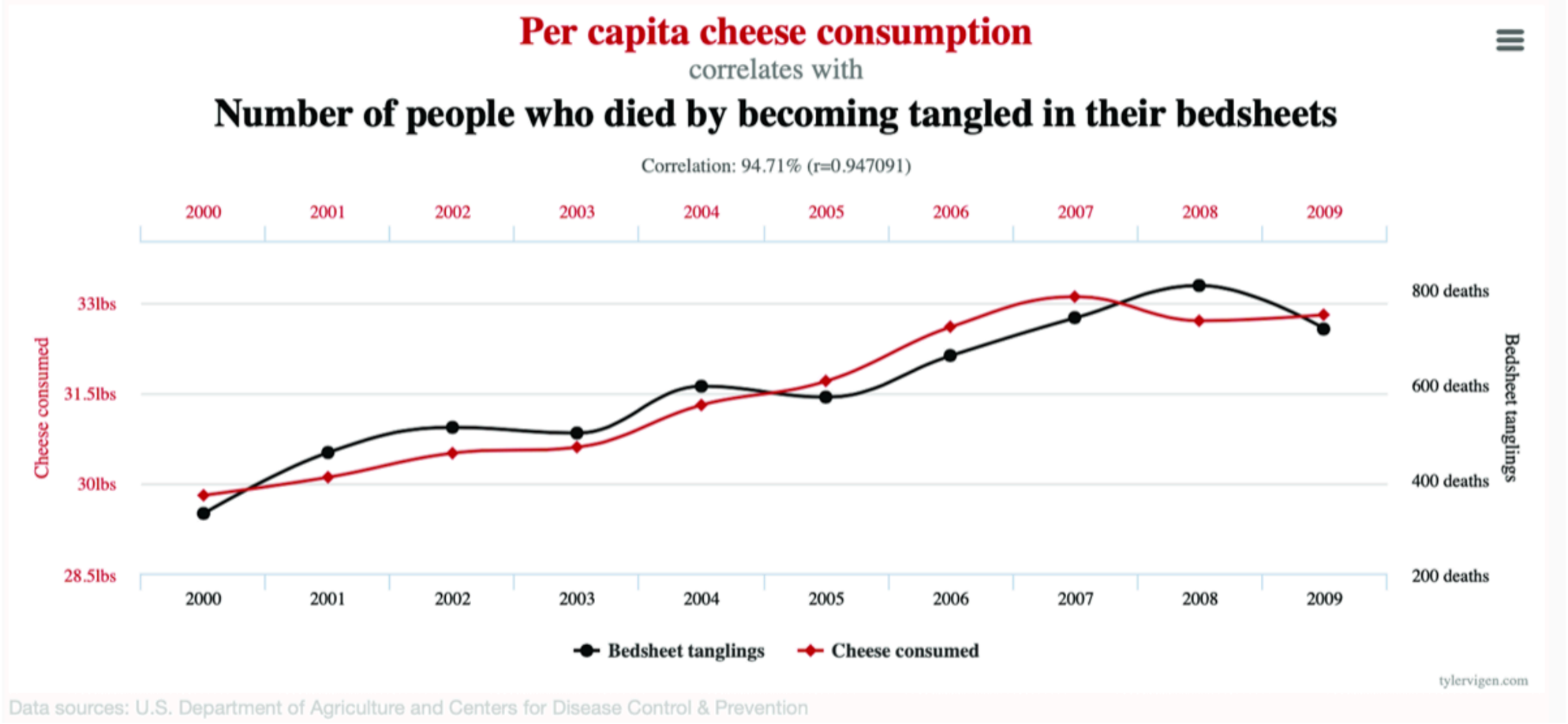


Problem 3: What is Lost or Misinterpreted?



https://www.tylervigen.com/spurious-correlations

Problem 3: What is Lost or Misinterpreted?

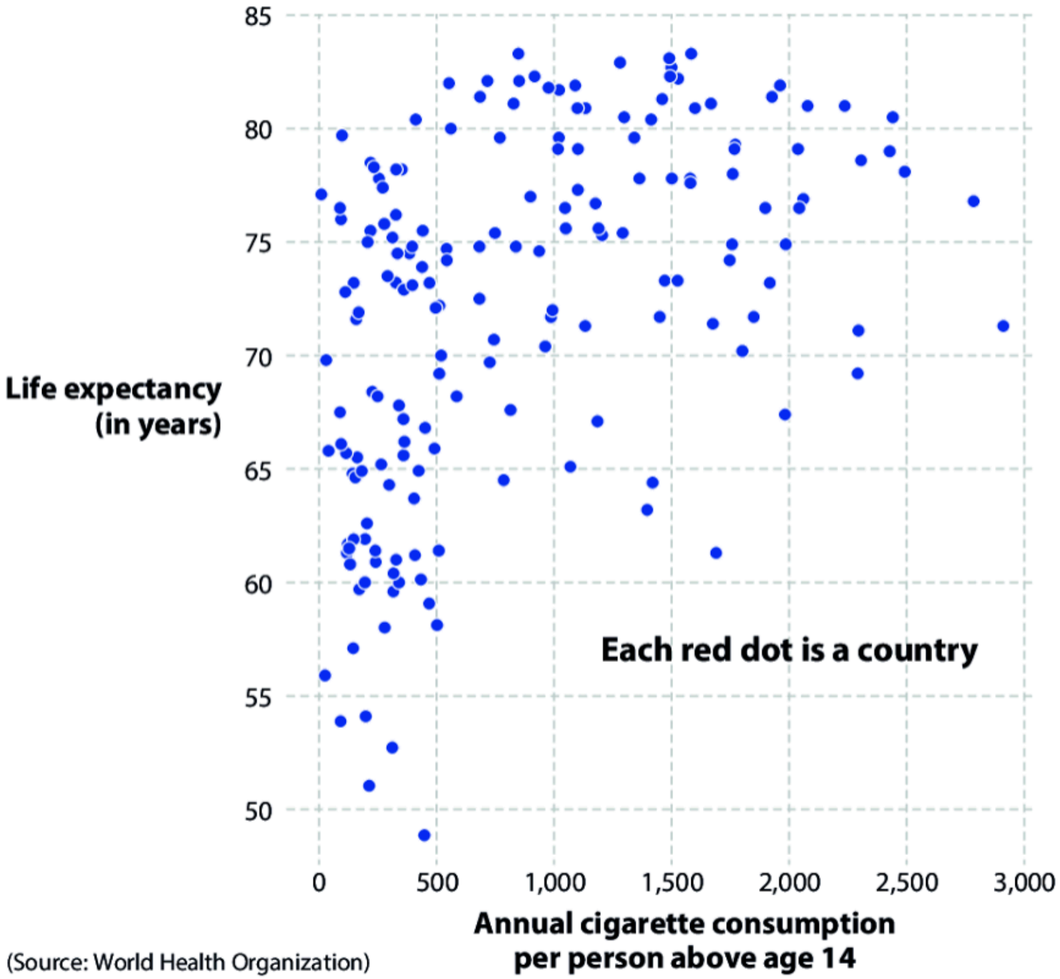


https://www.tylervigen.com/spurious-correlations

Problem 3: What is Lost or Misinterpreted?

~~“The more cigarettes we smoke, the longer we live!”~~

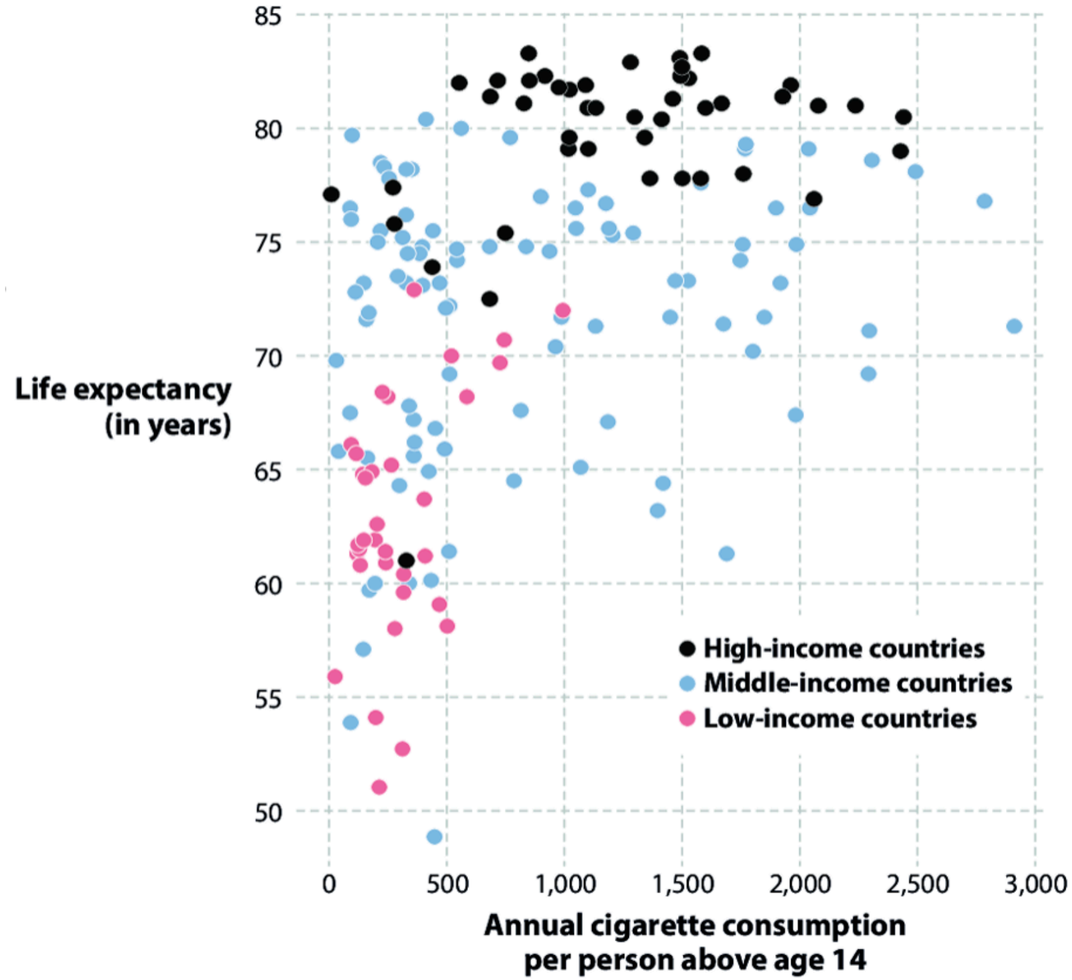
“There is a positive relationship between cigarette consumption and life expectancy at the country level”



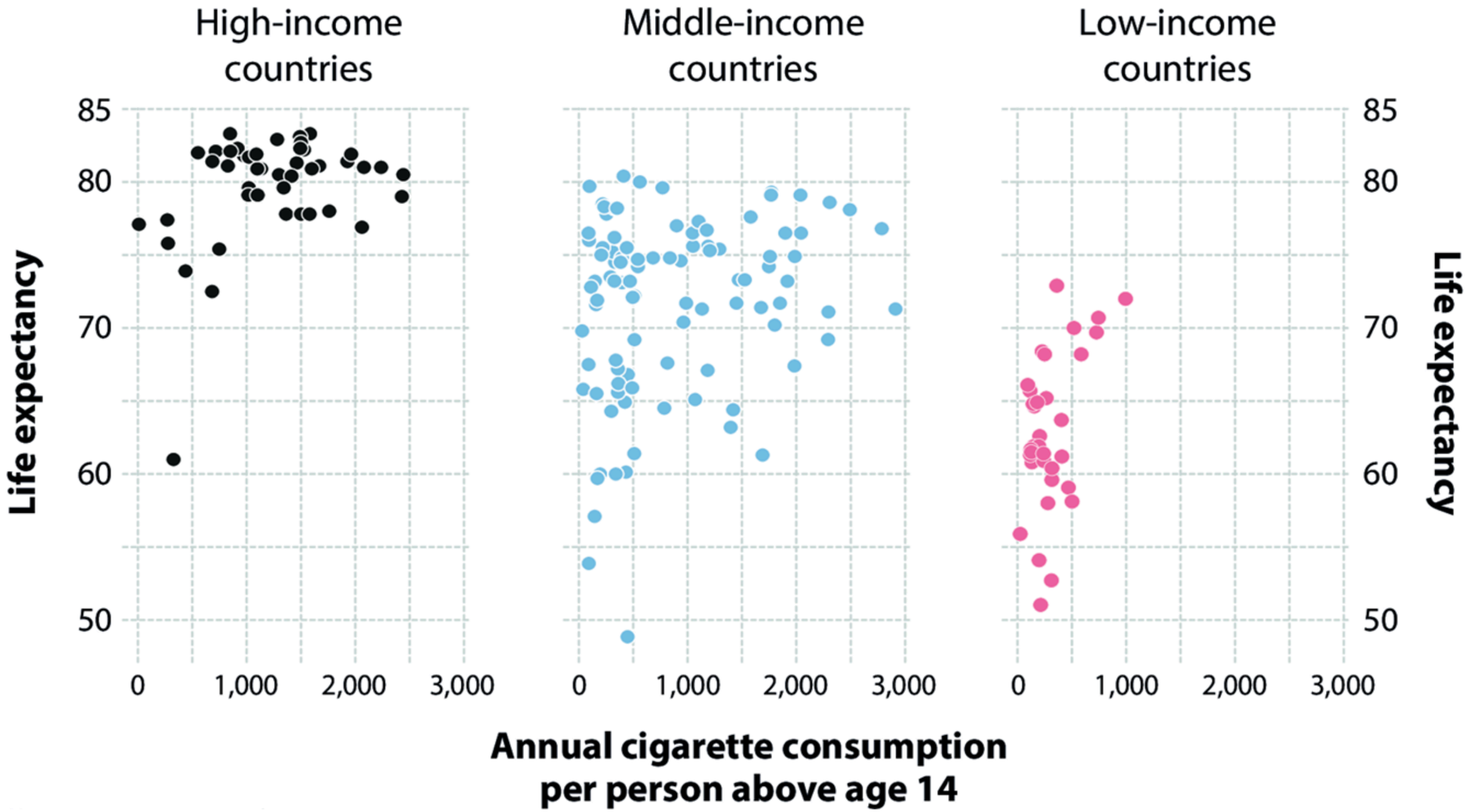
Problem 3: What is Lost or Misinterpreted?

~~“The more cigarettes we smoke, the longer we live!”~~

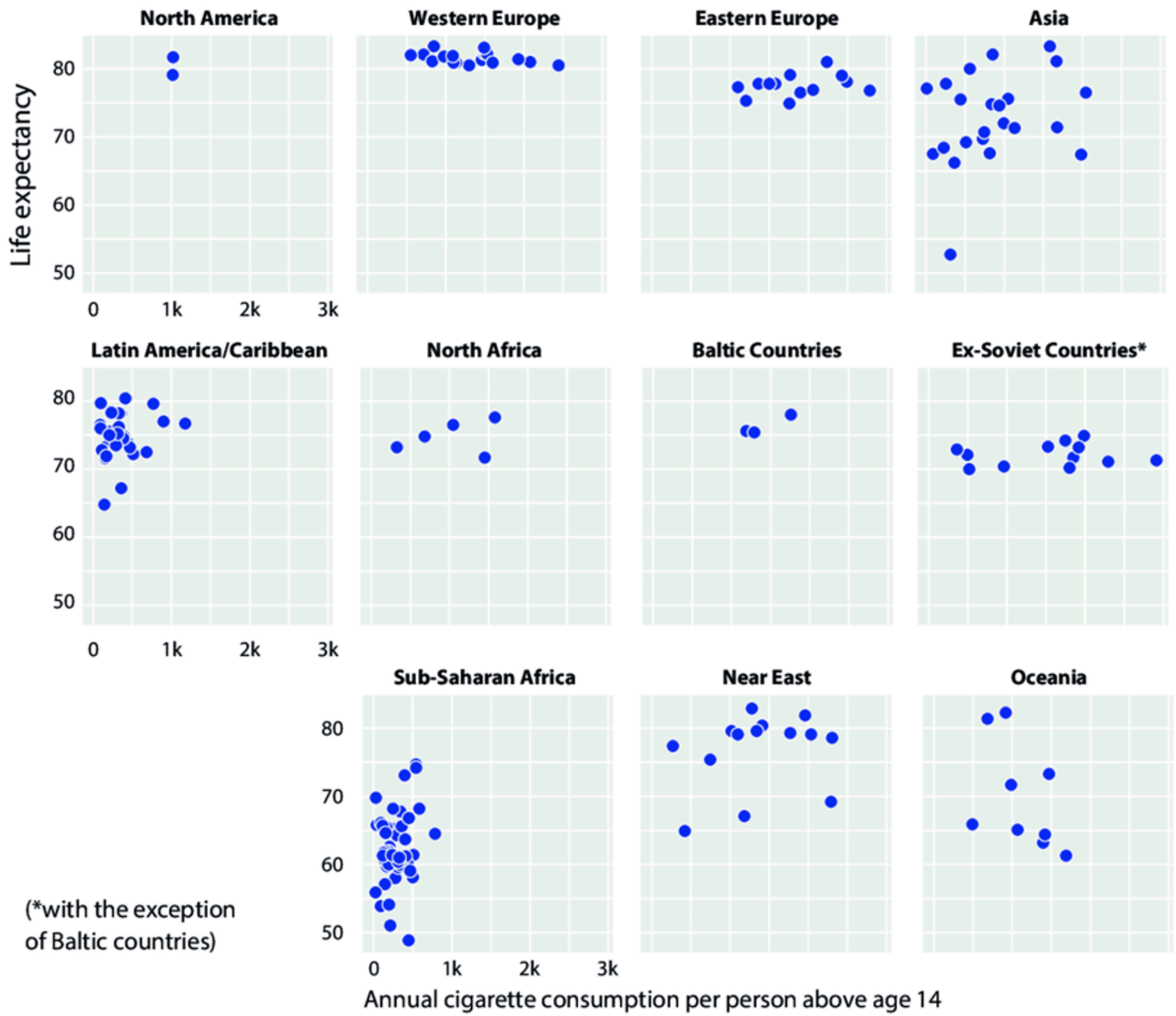
“There is a positive relationship between cigarette consumption and life expectancy at the country level”



Problem 3: What is Lost or Misinterpreted?



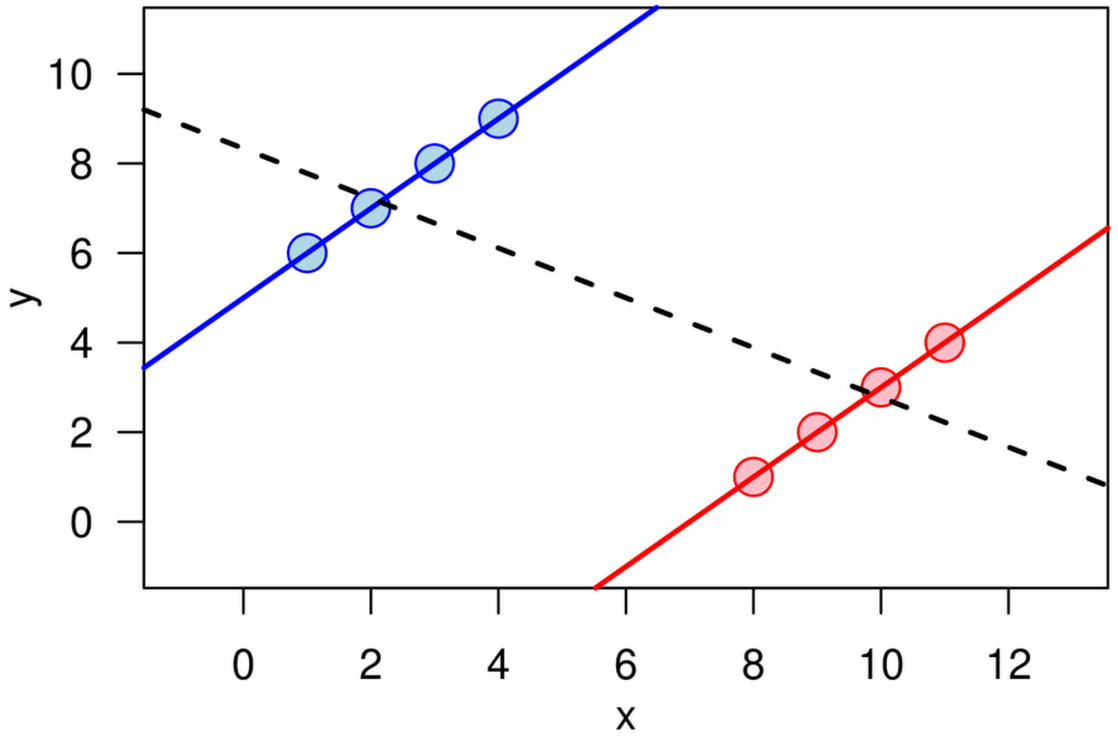
Problem 3: What is Lost or Misinterpreted?

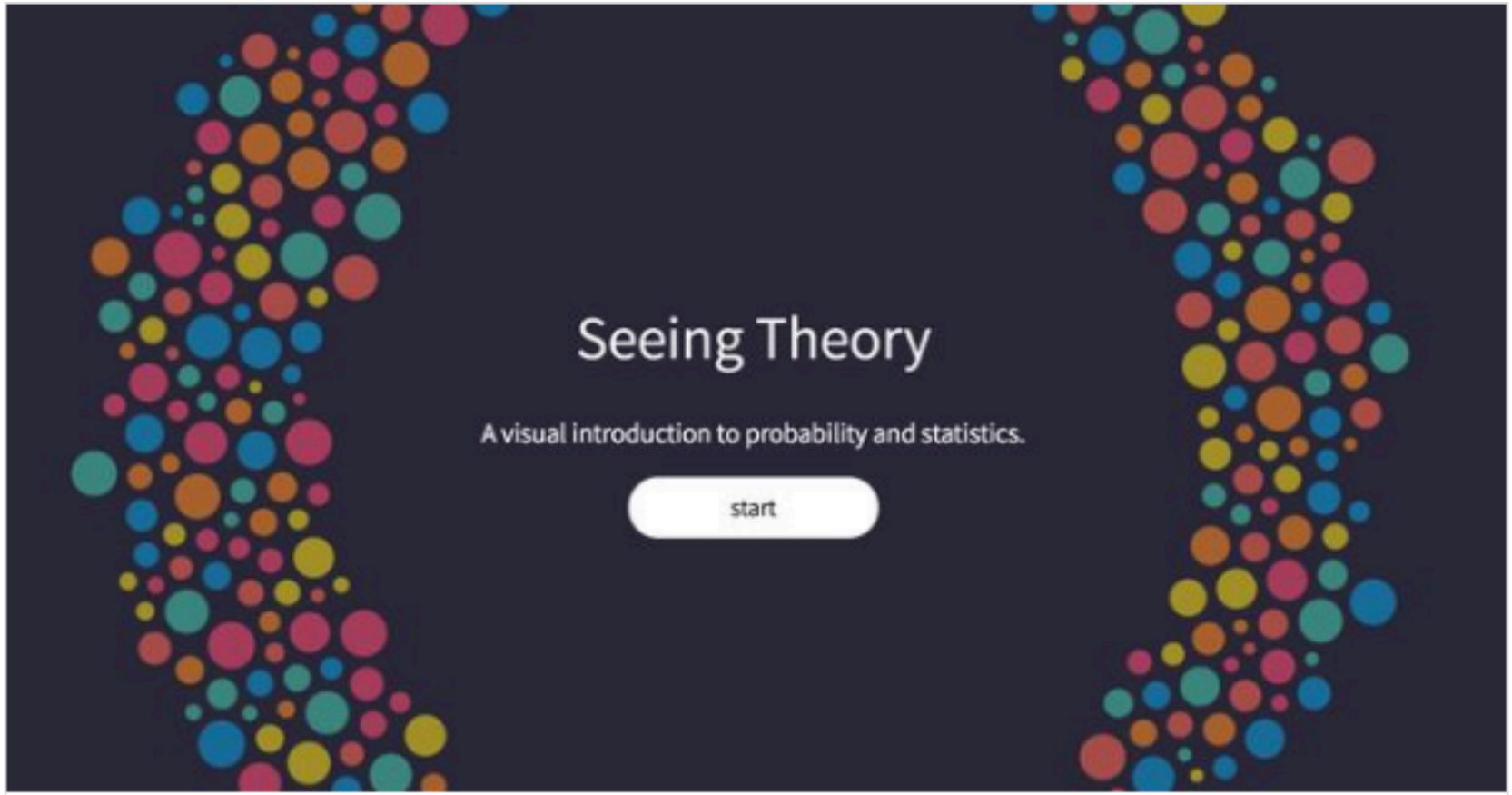


Problem 3: What is Lost or Misinterpreted?

Simpson's Paradox:

- trend that appears in several different groups of data but disappears or reverses when these groups are combined





Seeing Theory

A visual introduction to probability and statistics.

seeingtheory.io /

https://seeing-theory.brown.edu/

FIN

Upcoming Dates

Apr 25: Homework 4 (Due at 11:59pm Eastern)

Apr 24: Quiz 6 Released (all quizzes due May 2)

Apr 29: Final Group Activity

May 2:

- Homework 5 (Due at 11:59pm Eastern)
- Project Screencast Submission

May 12: Final Project Submission