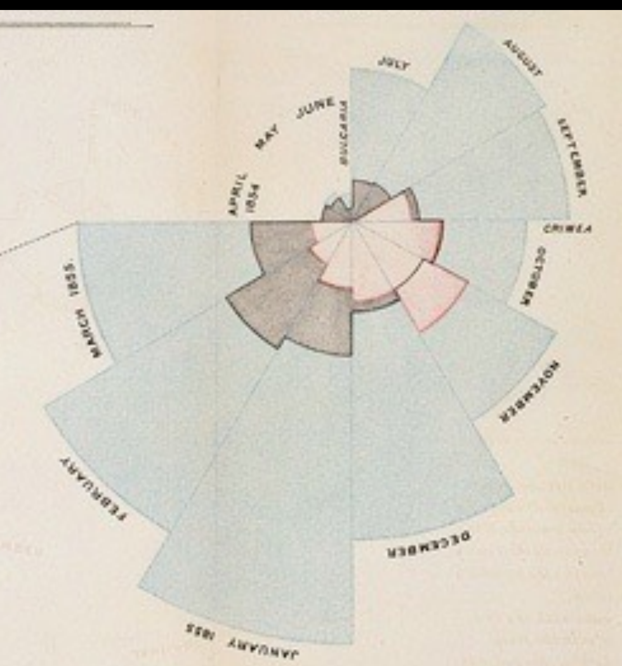
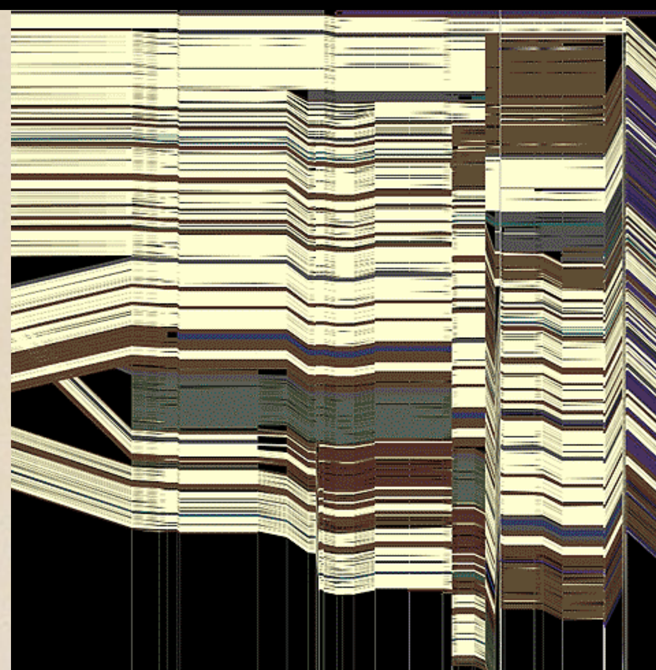
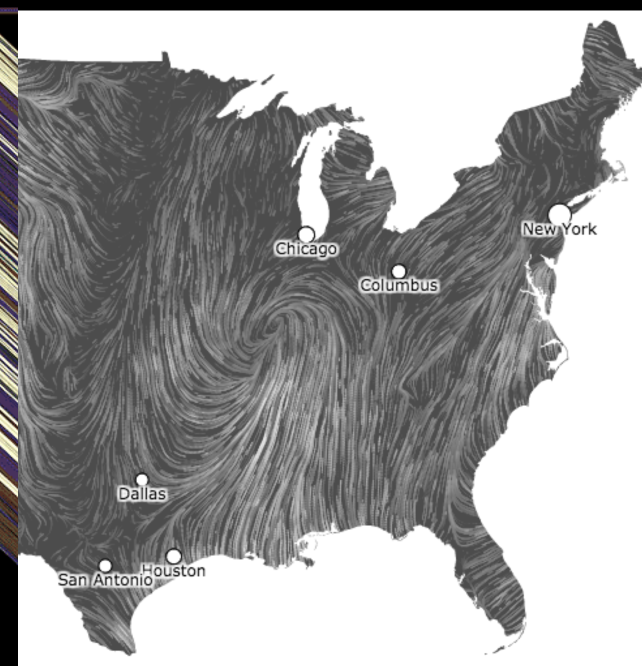
CSE 442 - Data Visualization

Visual Encoding Design







Jeffrey Heer University of Washington

A Design Space of Visual Encodings

Mapping Data to Visual Variables

Assign **data fields** (e.g., with *N, O, Q* types) to **visual channels** (*x, y, color, shape, size, …*) for a chosen **graphical mark** type (*point, bar, line, …*).

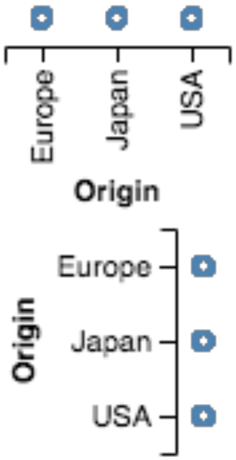
Additional concerns include choosing appropriate **encoding parameters** (*log scale, sorting, …*) and **data transformations** (*bin*, *group, aggregate, …*).

These options define a large combinatorial space, containing both useful and questionable charts!

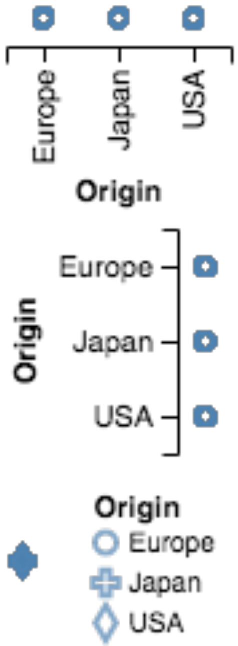
1D: Nominal

1D: Nominal

1D: Nominal

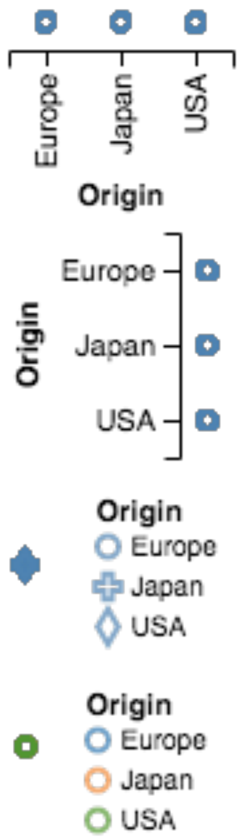


1D: Nominal

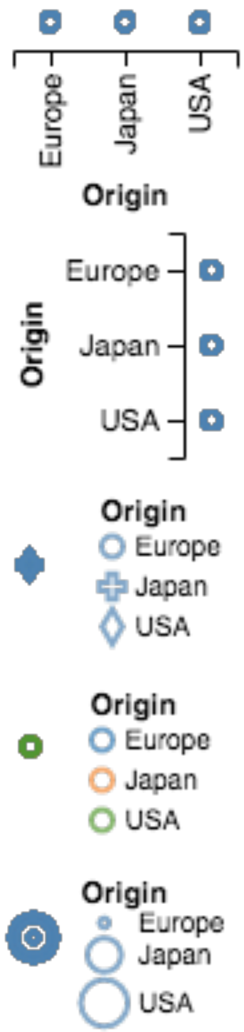


Europe
Japan
USA
Origin
Origin
Europe
Japan
USA
Origin
Europe
Japan
USA

1D: Nominal

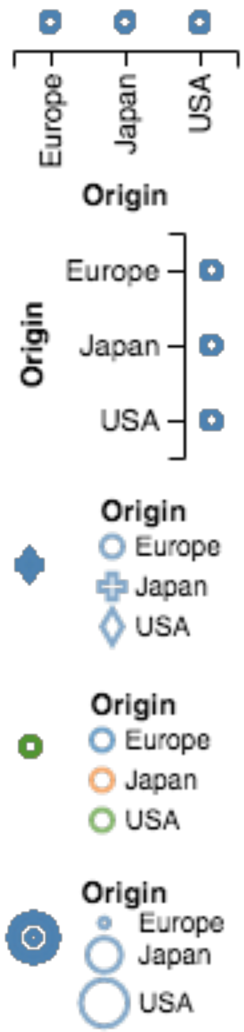

Europe
Japan
USA
Origin
Origin
Europe
Japan
USA
Origin
Europe
Japan
USA
Origin
Europe
Japan
USA

1D: Nominal


Europe
Japan
USA
Origin
Origin
Europe
Japan
USA
Origin
Europe
Japan
USA
Origin
Europe
Japan
USA
Origin
Europe
Japan
USA

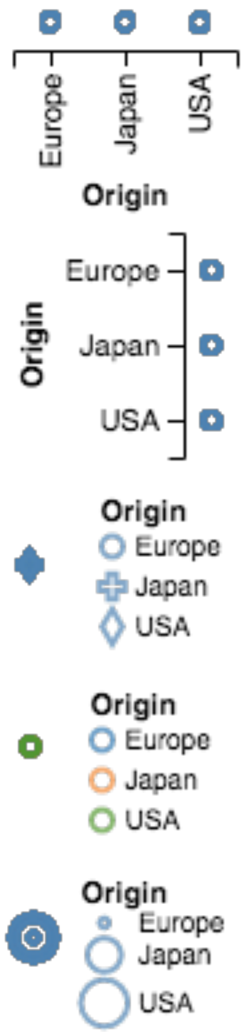
1D: Nominal

Raw



1D: Nominal

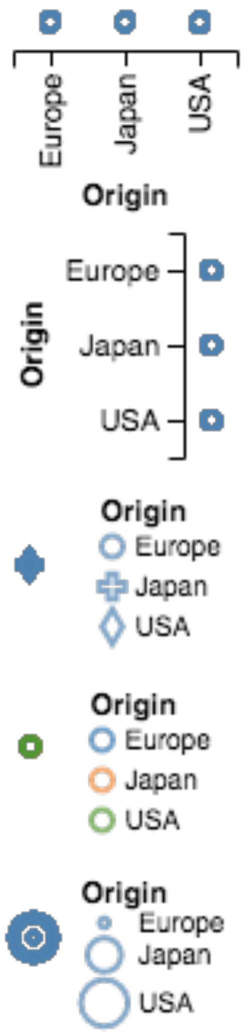
Raw



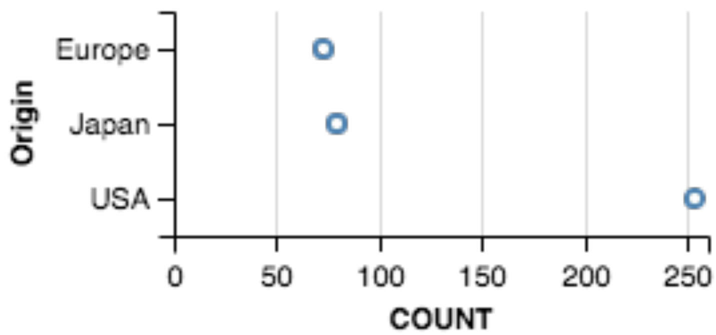
Aggregate (Count)

1D: Nominal

Raw

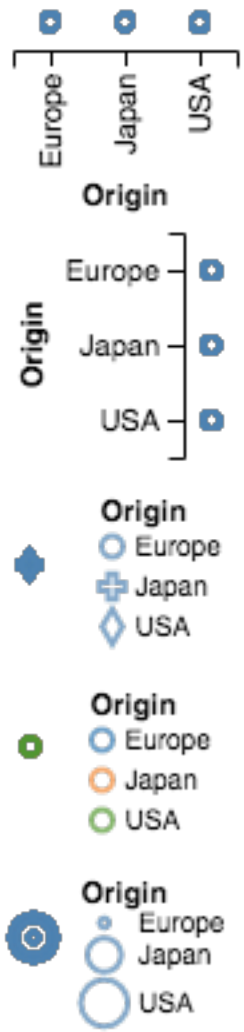


Aggregate (Count)

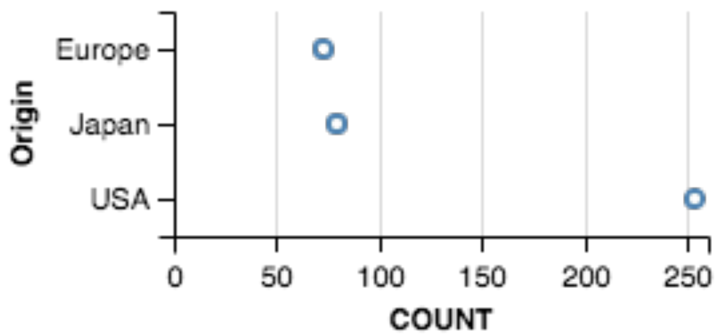


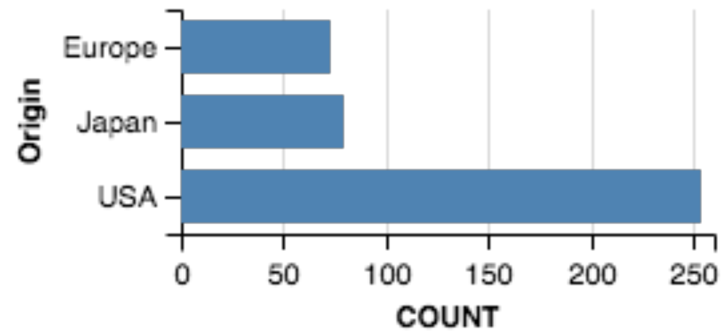
1D: Nominal

Raw



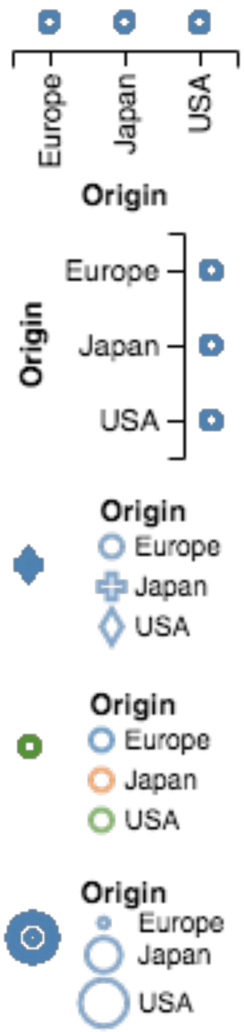
Aggregate (Count)



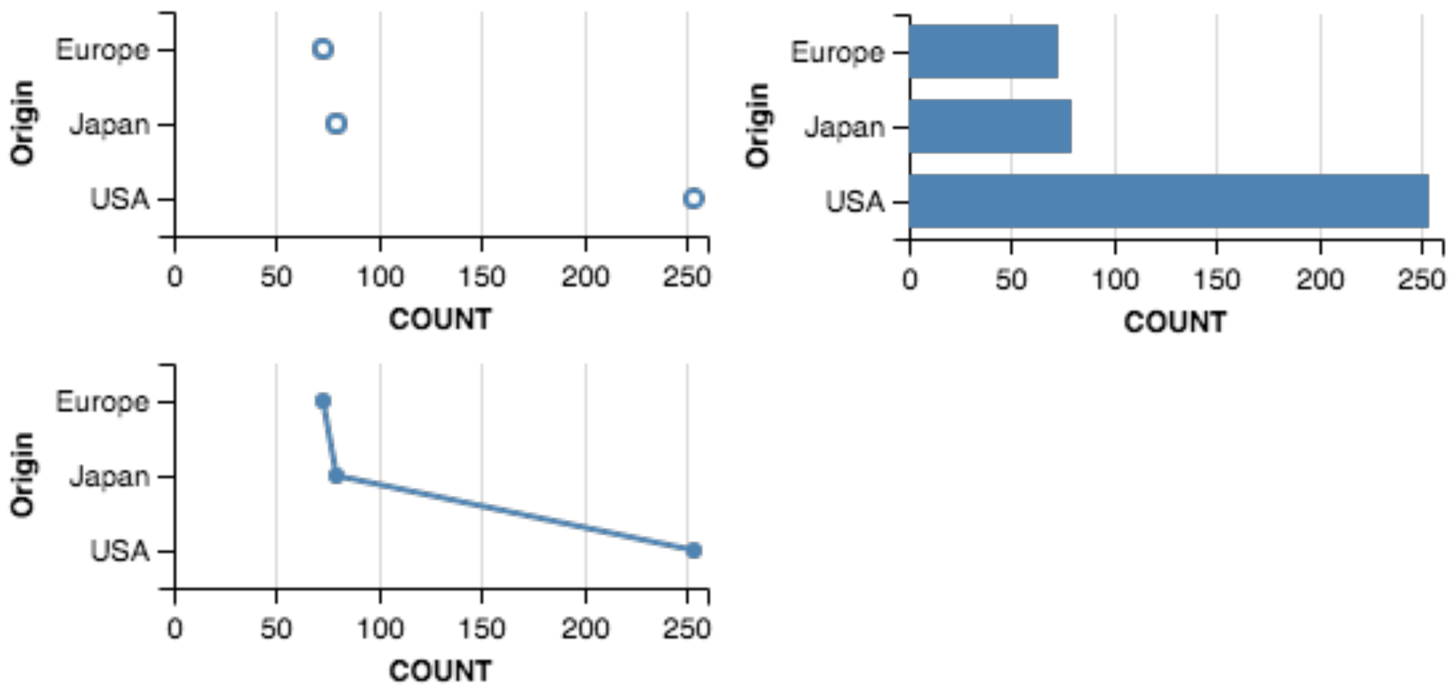


1D: Nominal

Raw

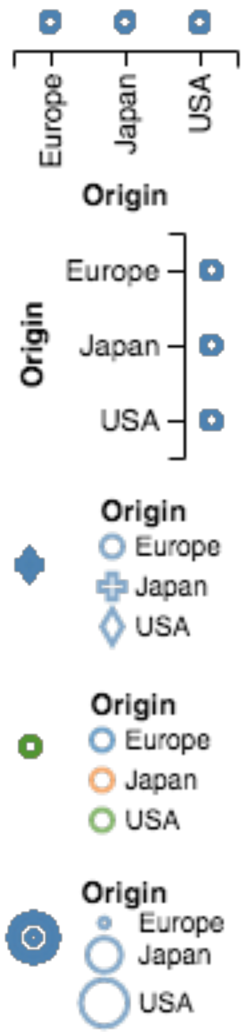


Aggregate (Count)

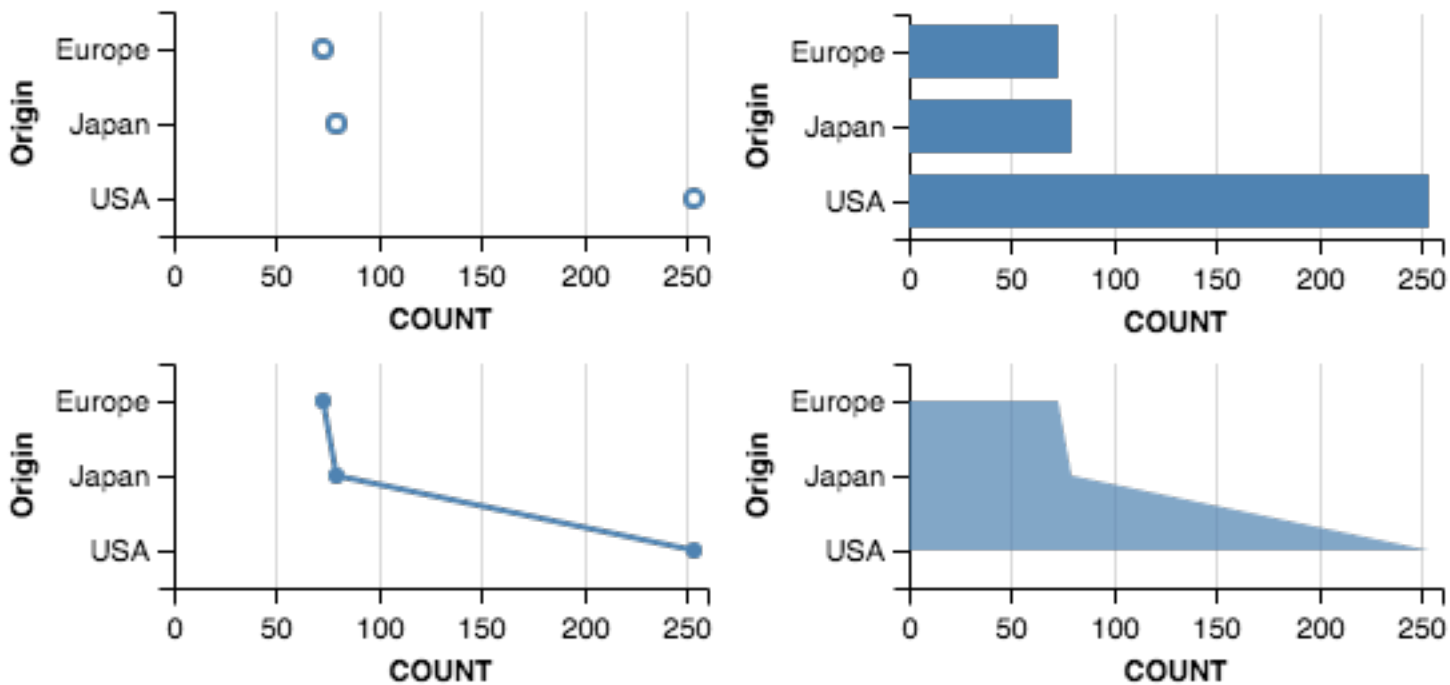


1D: Nominal

Raw

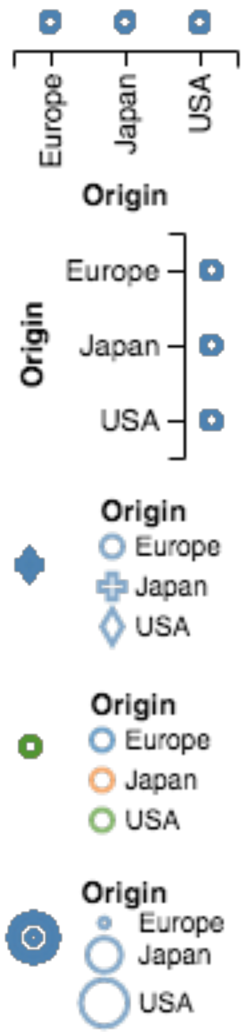


Aggregate (Count)

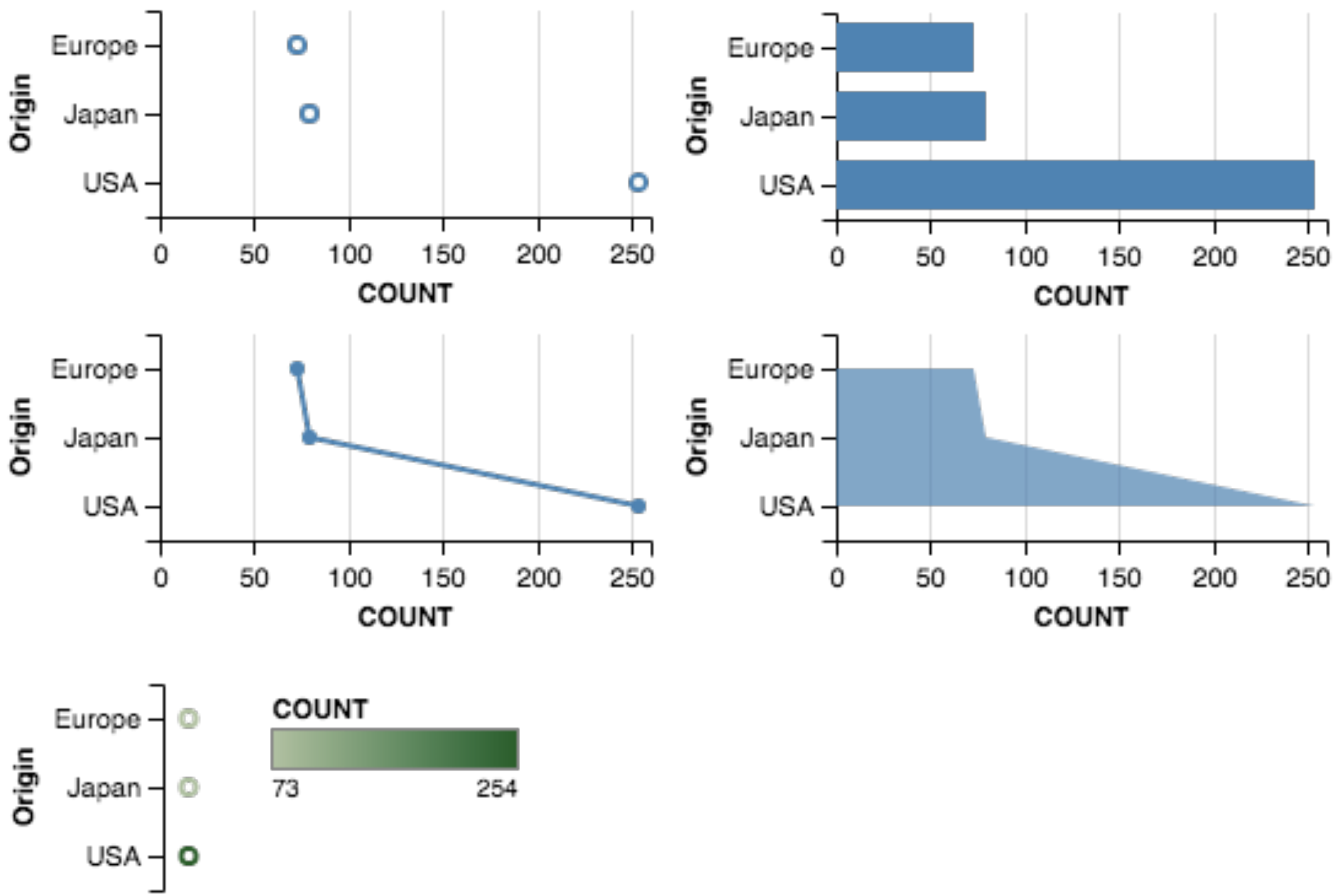


1D: Nominal

Raw



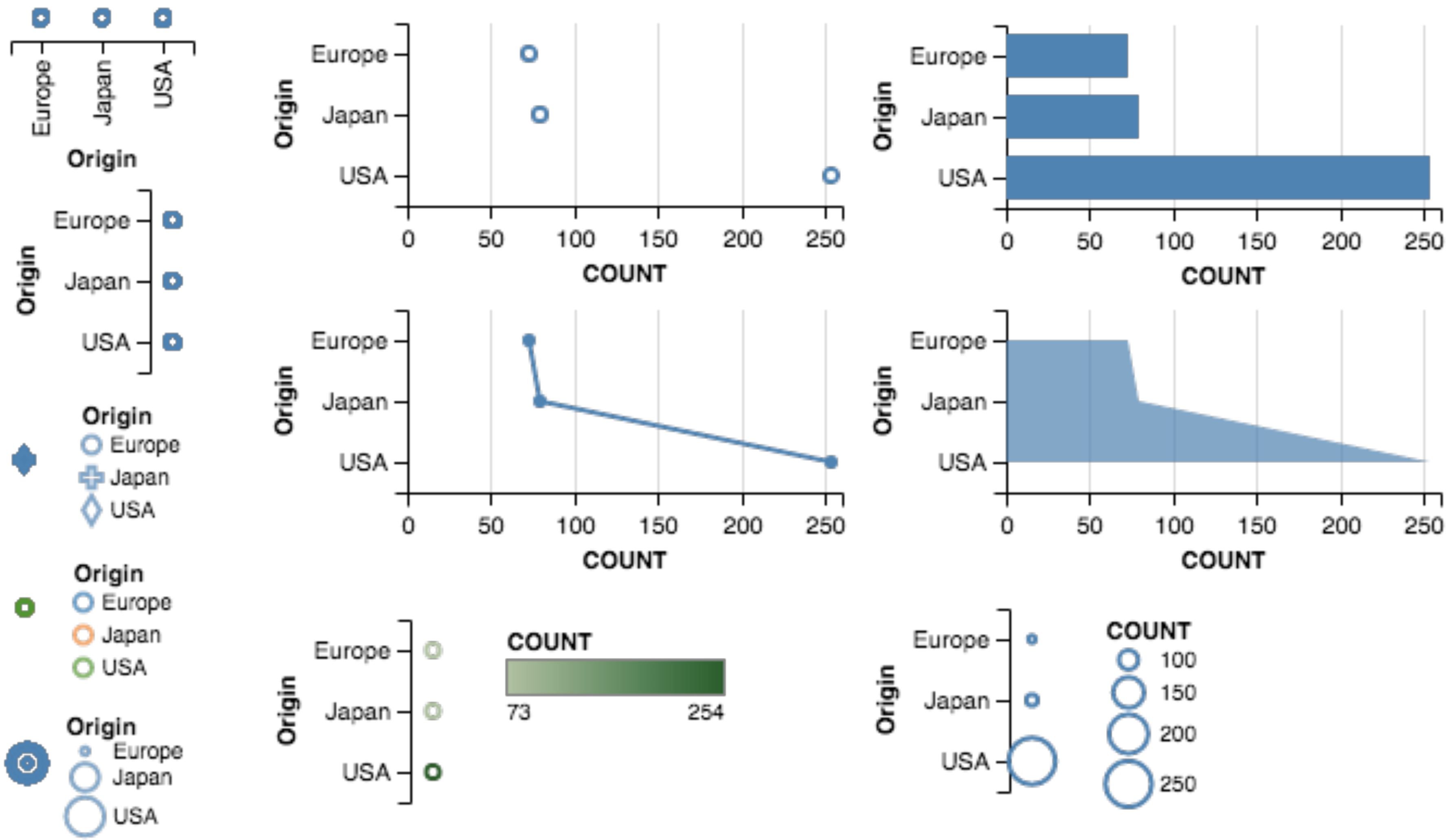
Aggregate (Count)



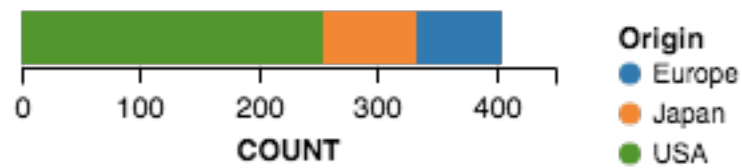
1D: Nominal

Raw

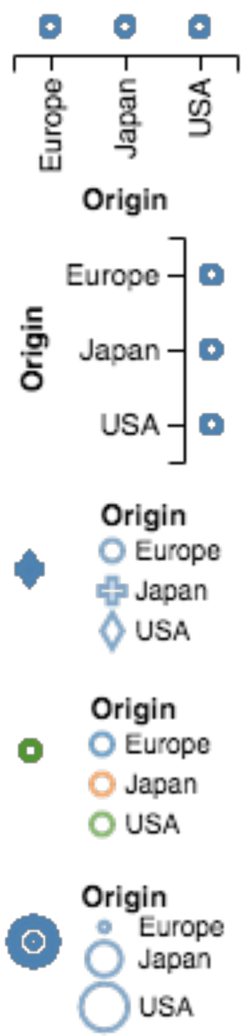
Aggregate (Count)



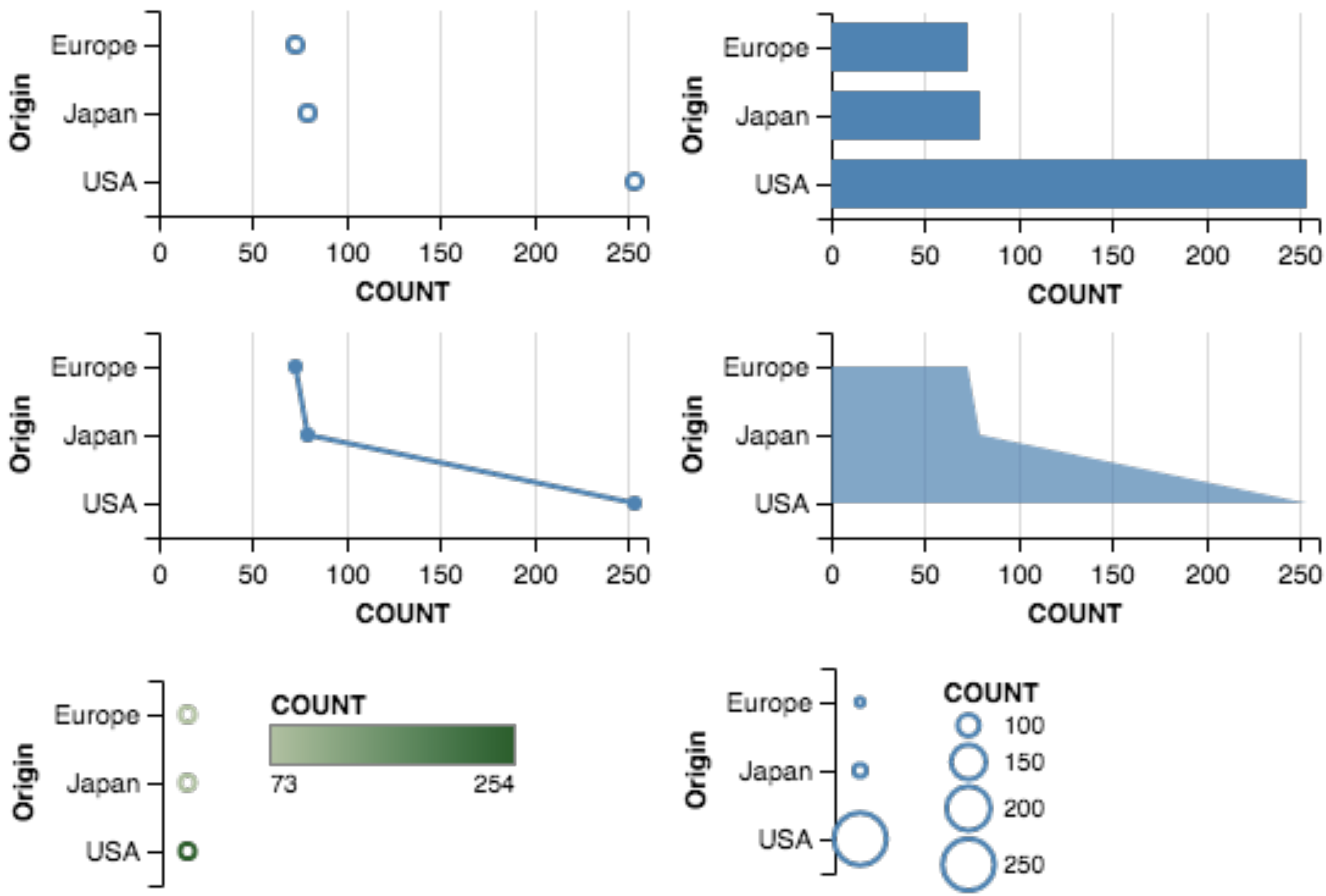
1D: Nominal



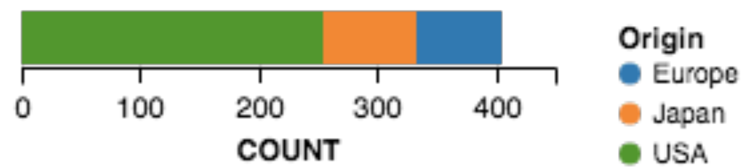
Raw



Aggregate (Count)

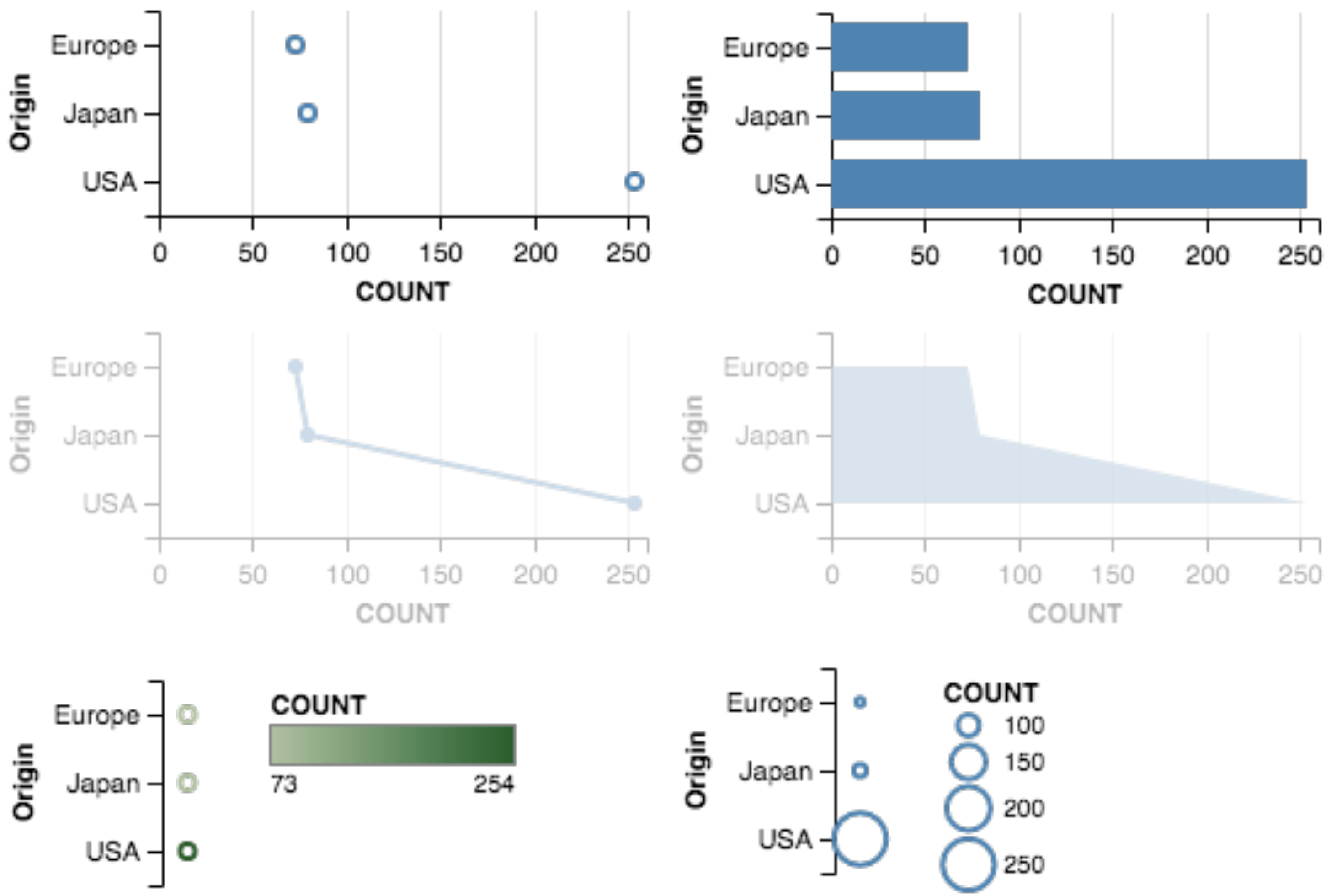


Expressive?



Raw

Aggregate (Count)



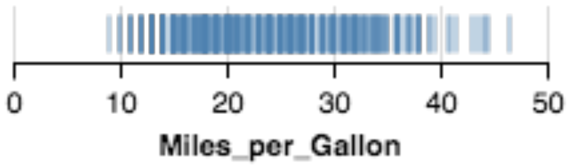
1D: Quantitative

1D: Quantitative

Raw

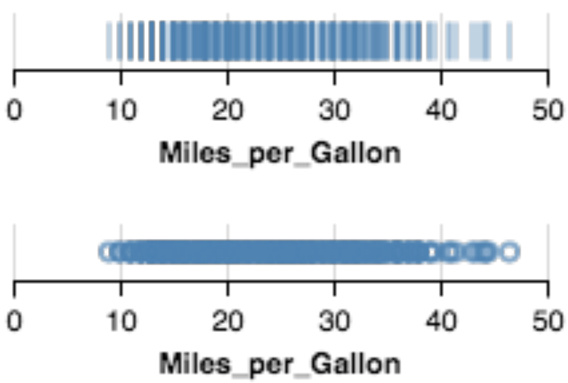
1D: Quantitative

Raw



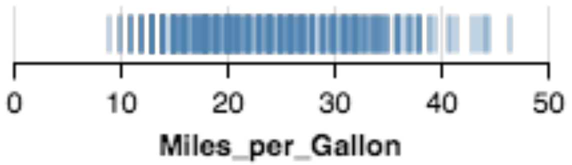
1D: Quantitative

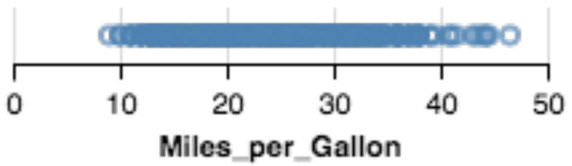
Raw

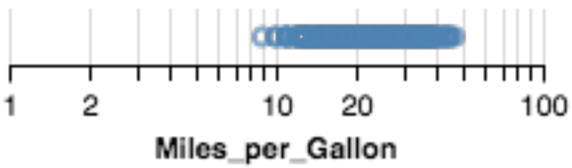


1D: Quantitative

Raw

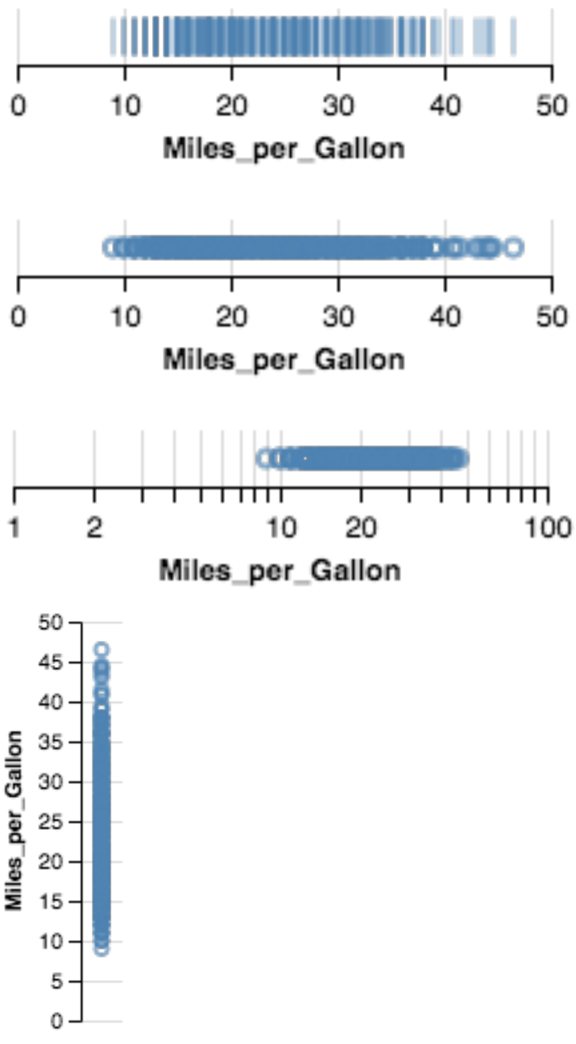






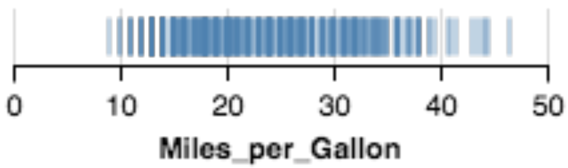
1D: Quantitative

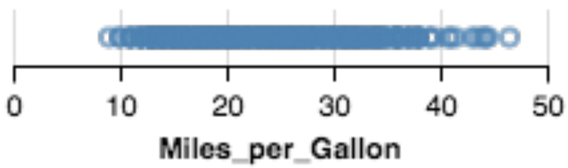
Raw

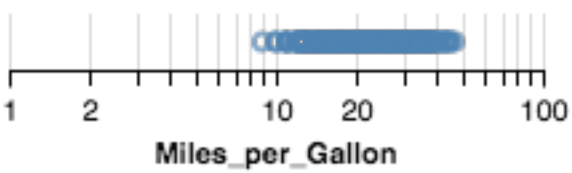


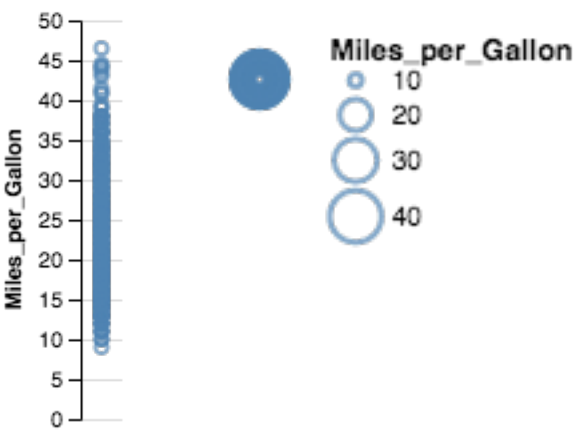
1D: Quantitative

Raw



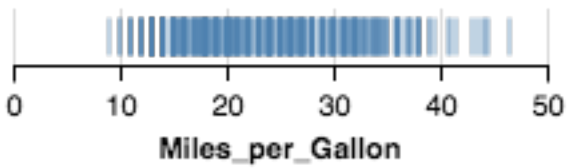


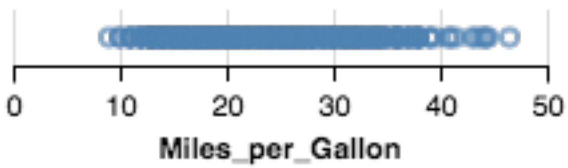


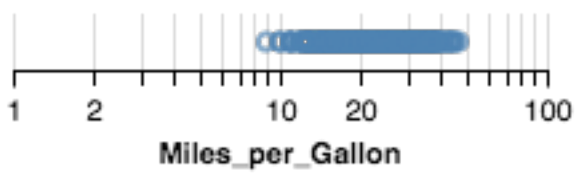


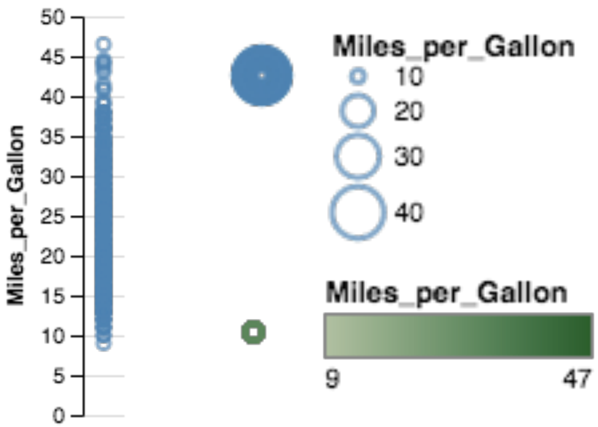
1D: Quantitative

Raw





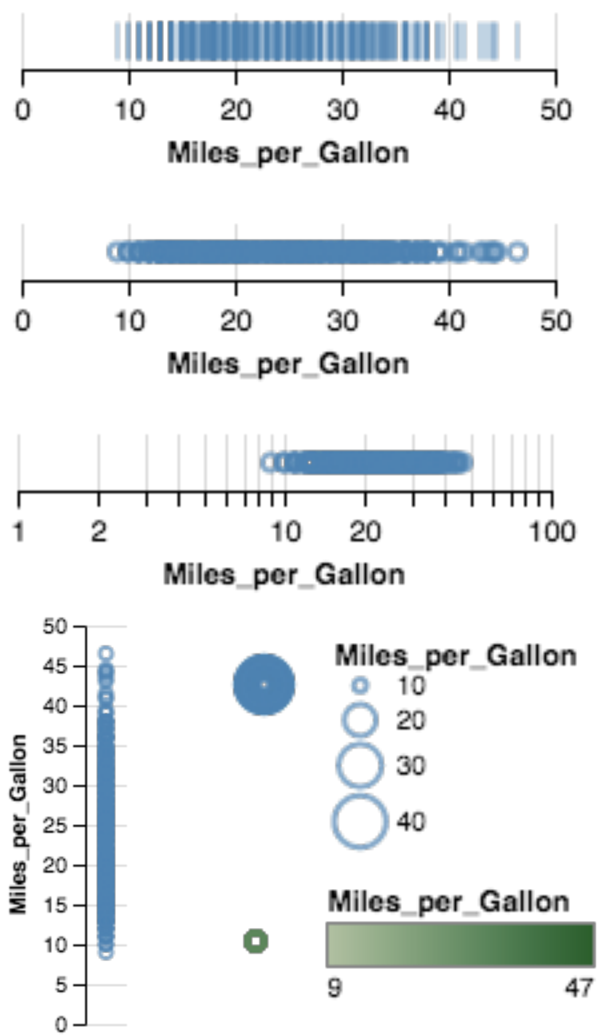




1D: Quantitative

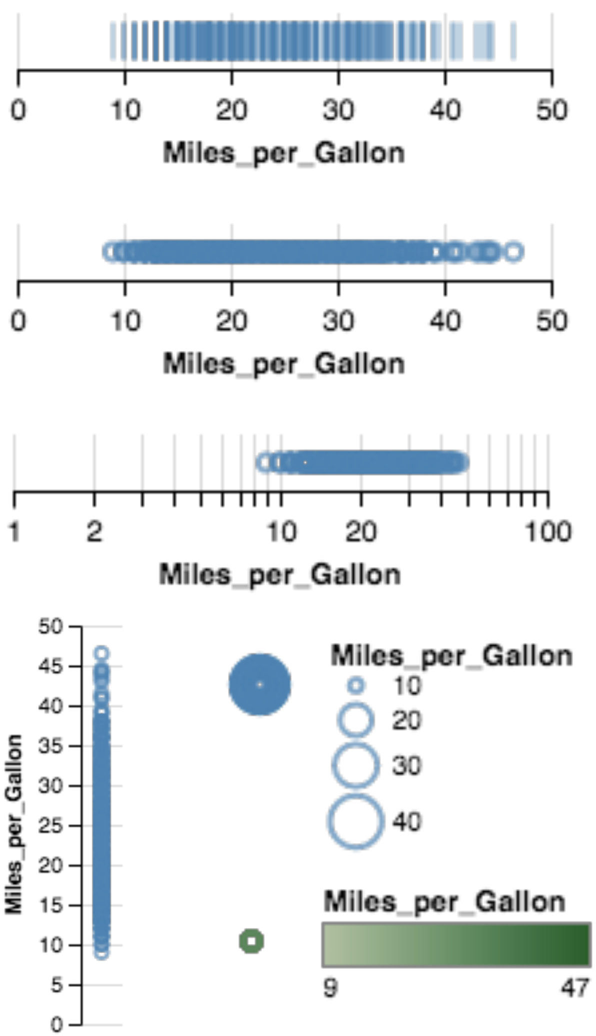
Raw

Aggregate (Count)

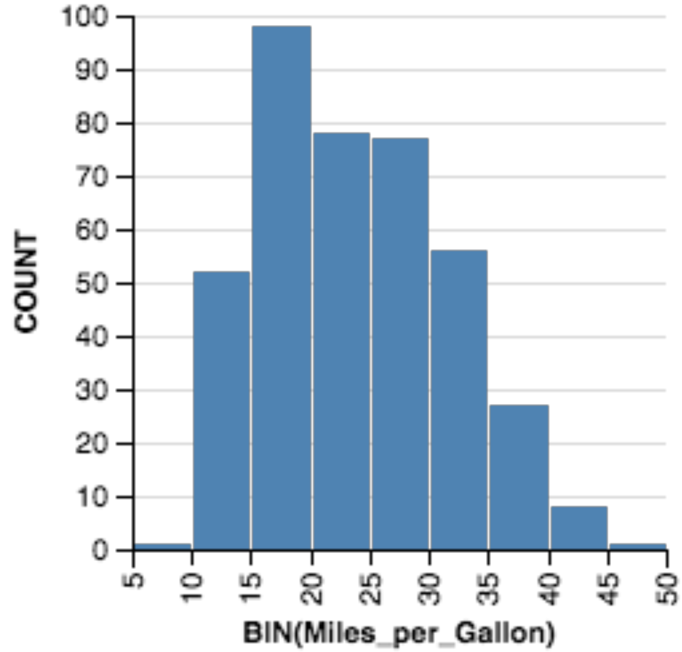


1D: Quantitative

Raw

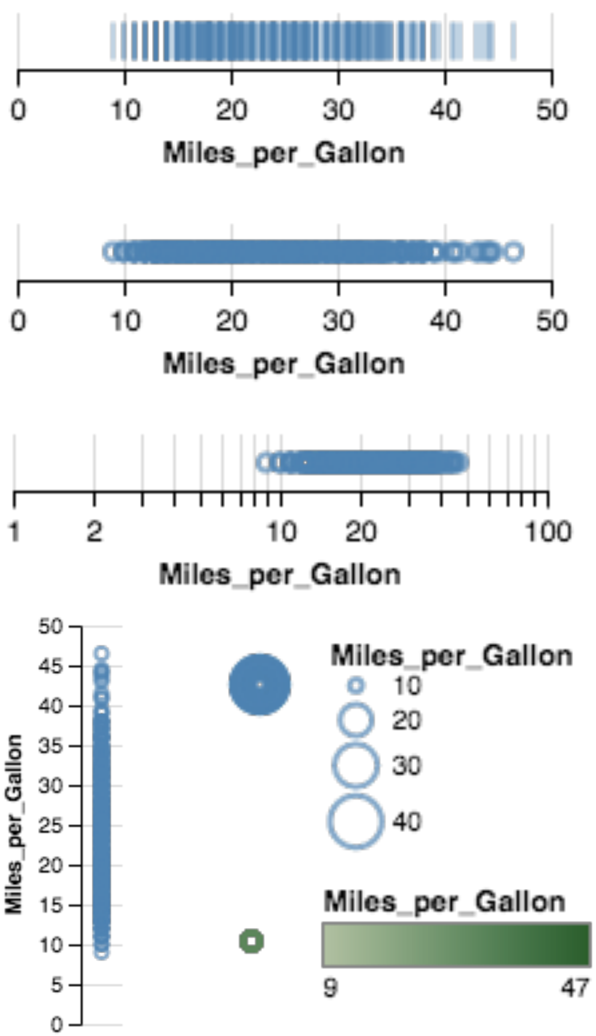


Aggregate (Count)

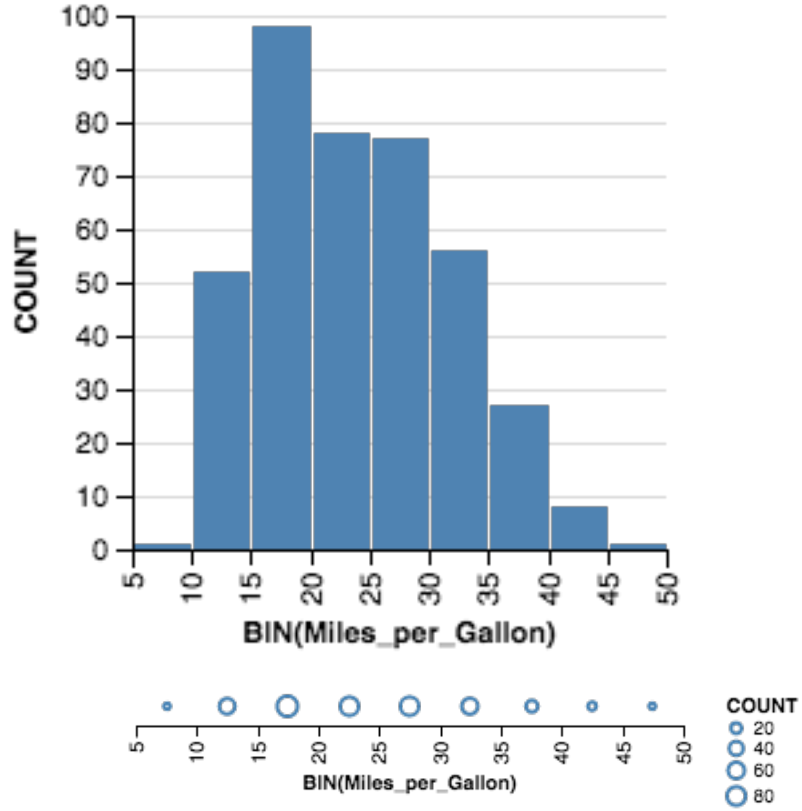


1D: Quantitative

Raw

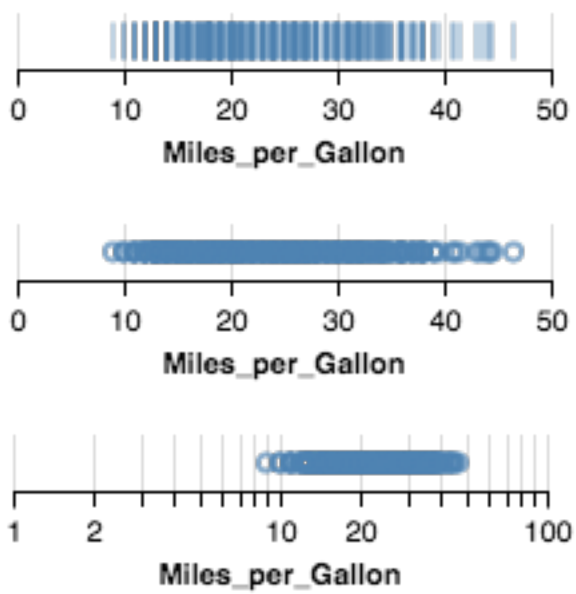


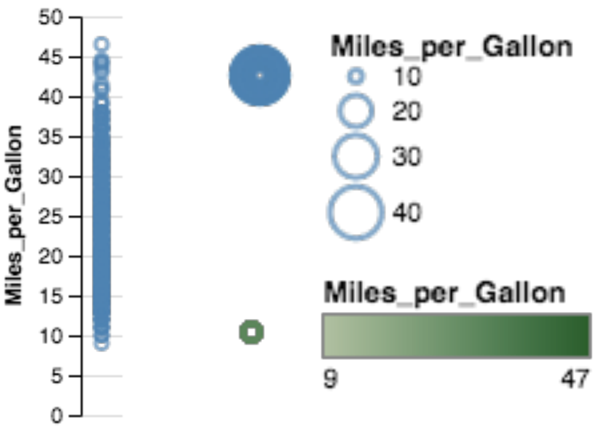
Aggregate (Count)



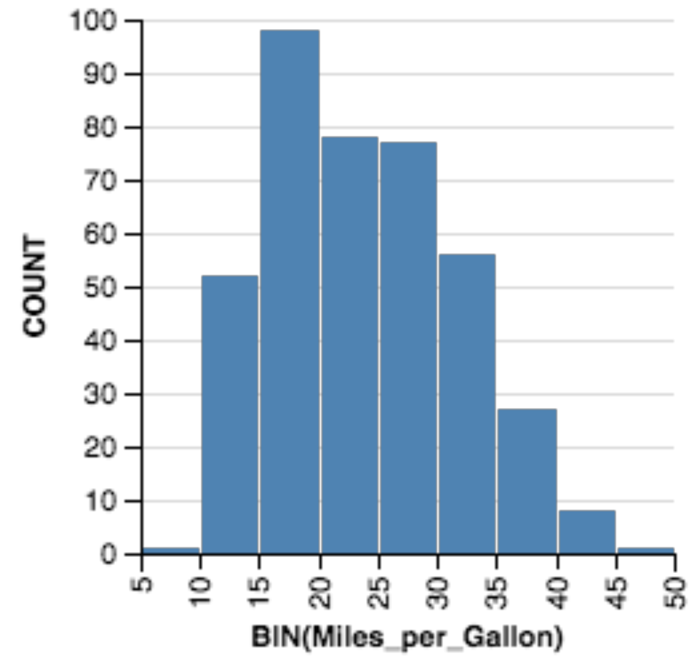
1D: Quantitative

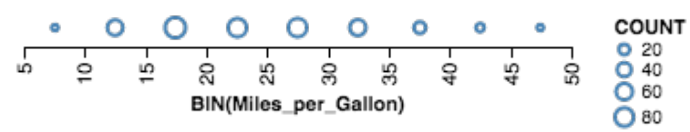
Raw

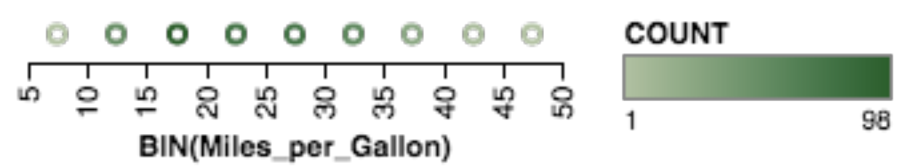




Aggregate (Count)

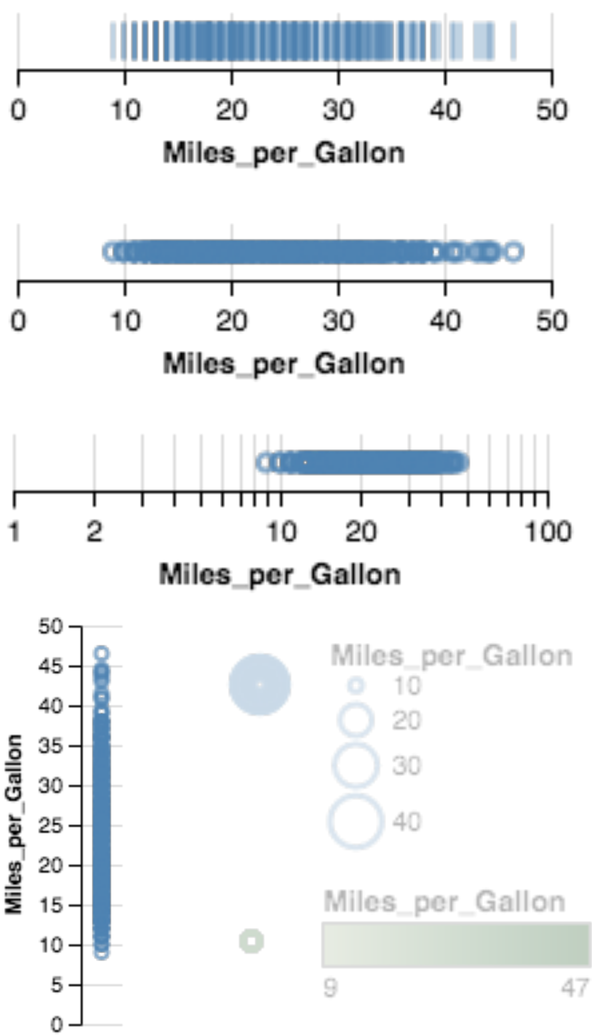




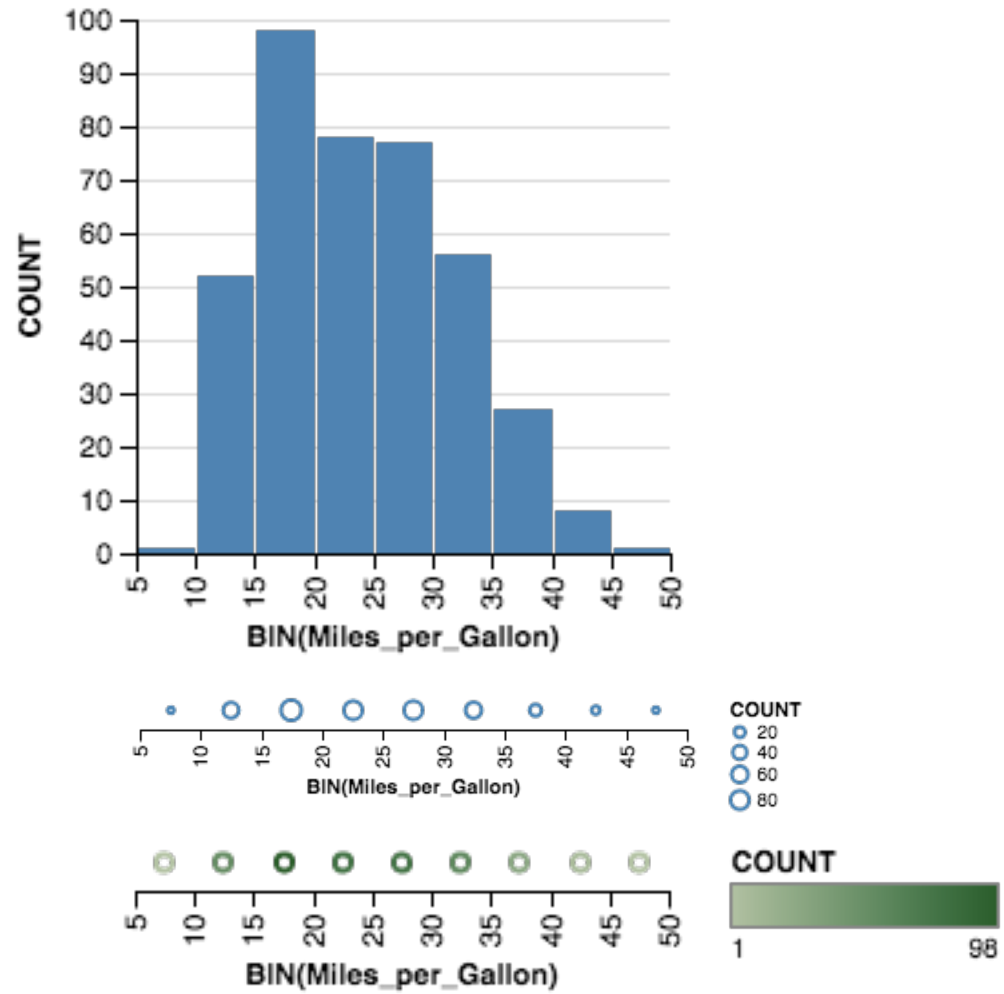


Expressive?

Raw

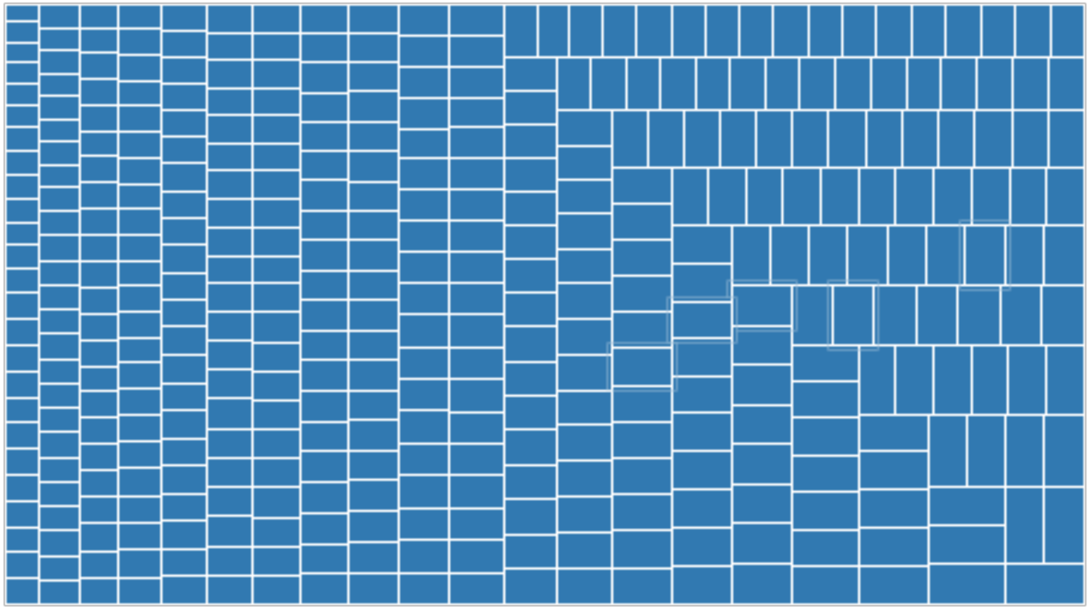


Aggregate (Count)



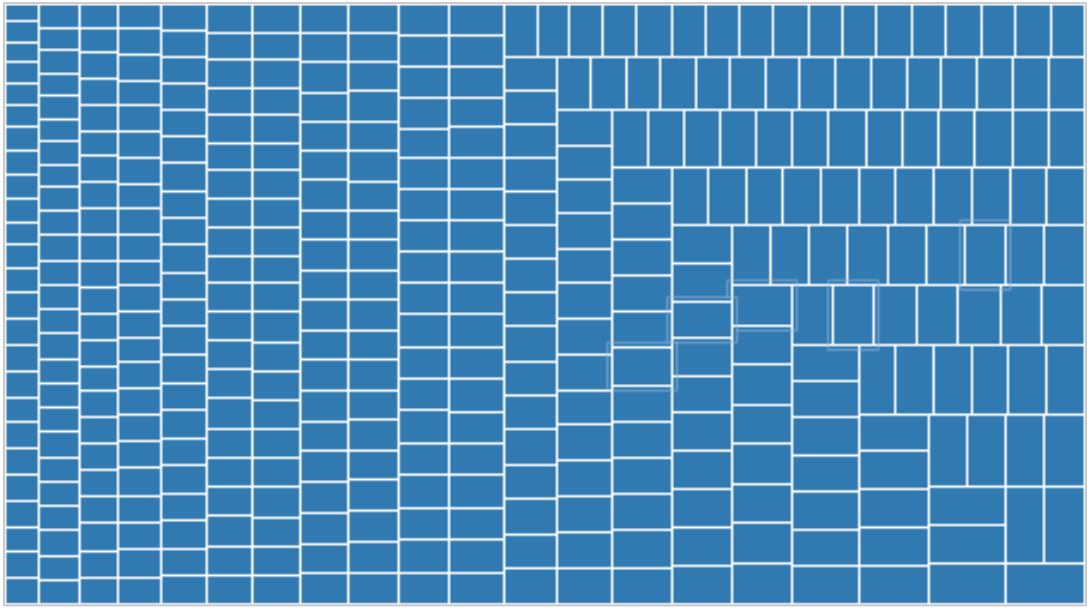
Raw (with Layout Algorithm)

Raw (with Layout Algorithm)

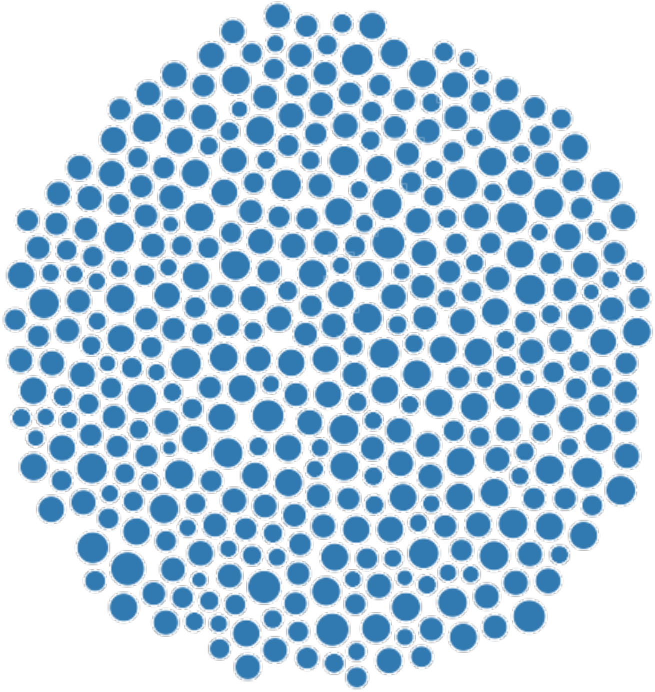


Treemap

Raw (with Layout Algorithm)

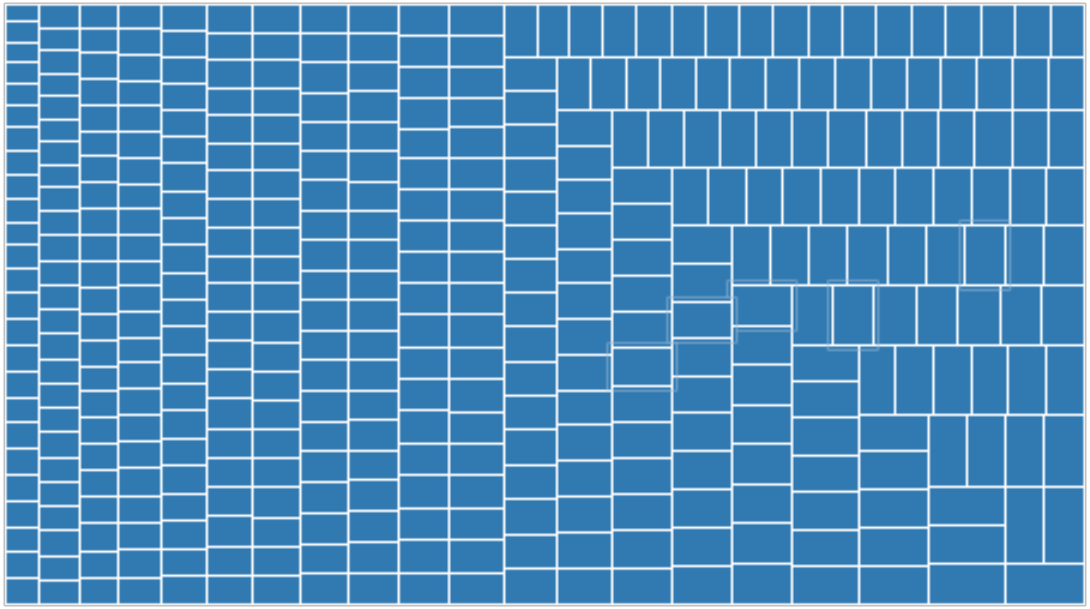


Treemap

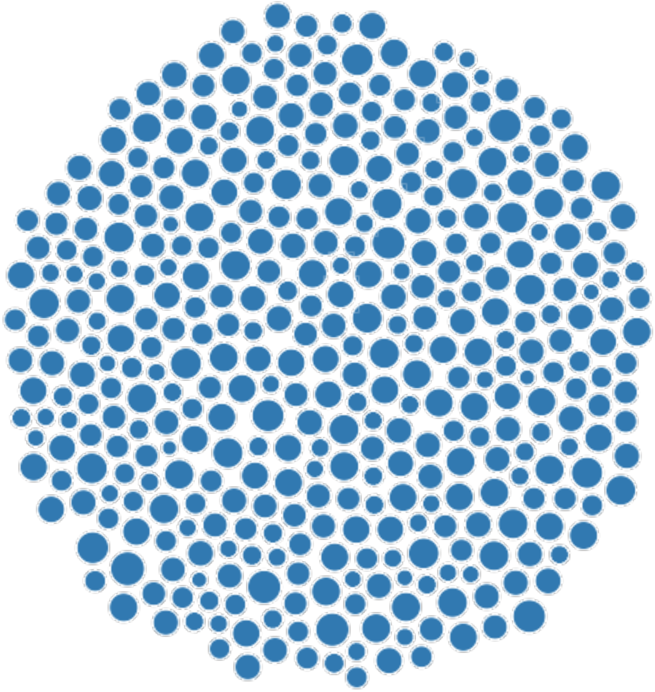


Bubble Chart

Raw (with Layout Algorithm)



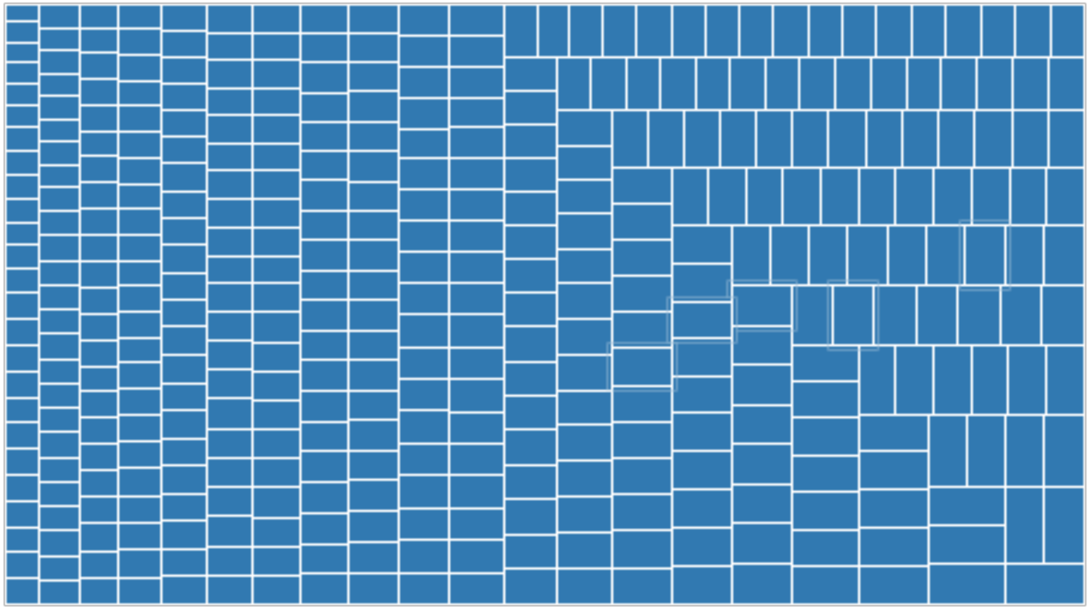
Treemap



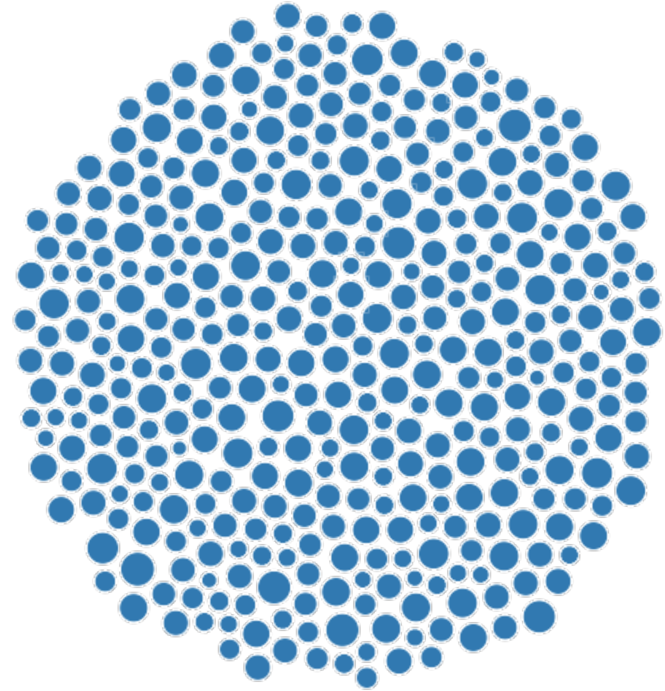
Bubble Chart

Aggregate (Distributions)

Raw (with Layout Algorithm)

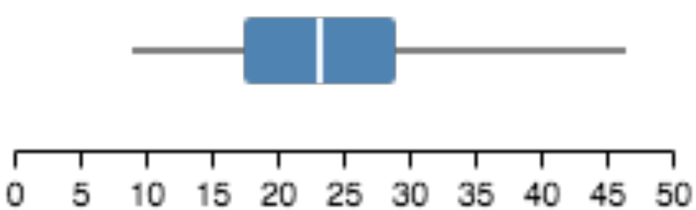


Treemap



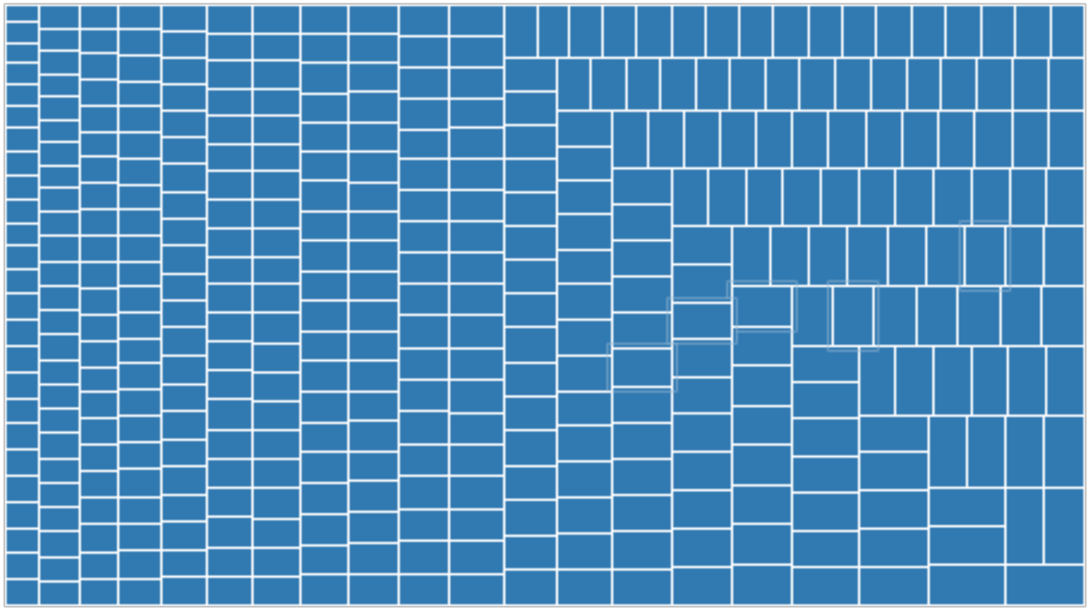
Bubble Chart

Aggregate (Distributions)

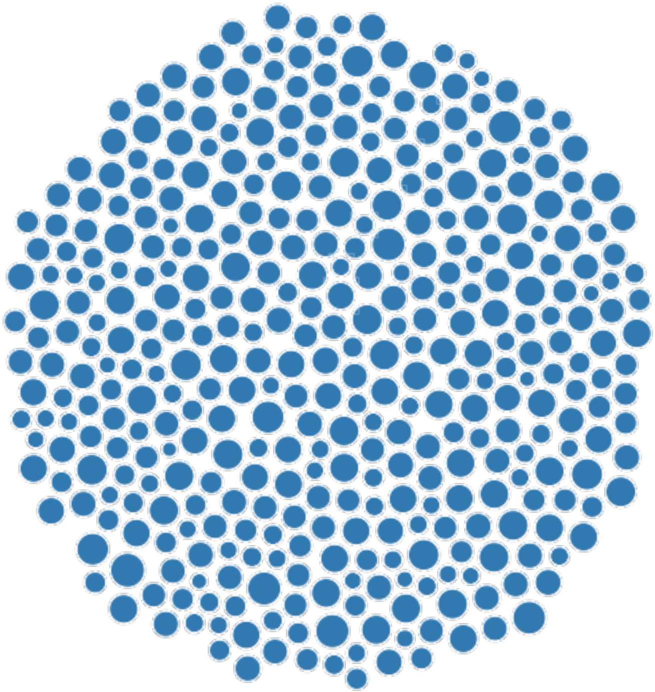


Box Plot

Raw (with Layout Algorithm)

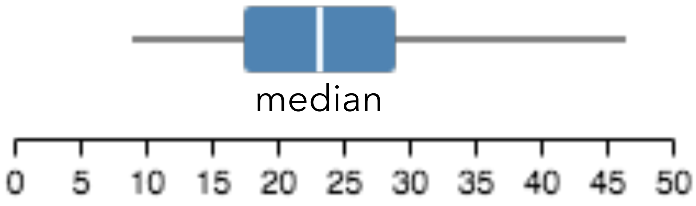


Treemap



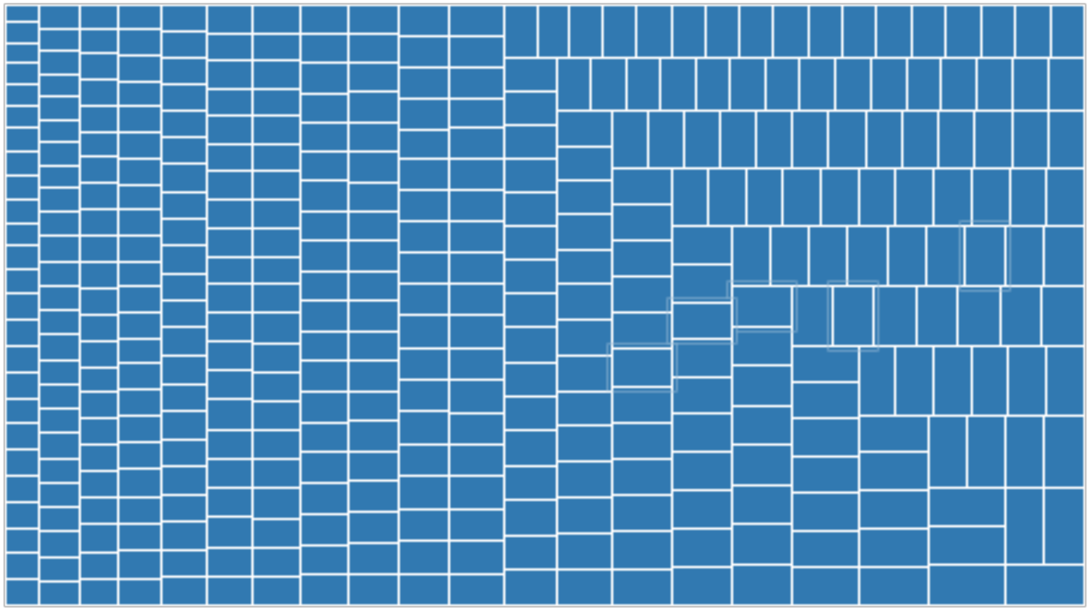
Bubble Chart

Aggregate (Distributions)

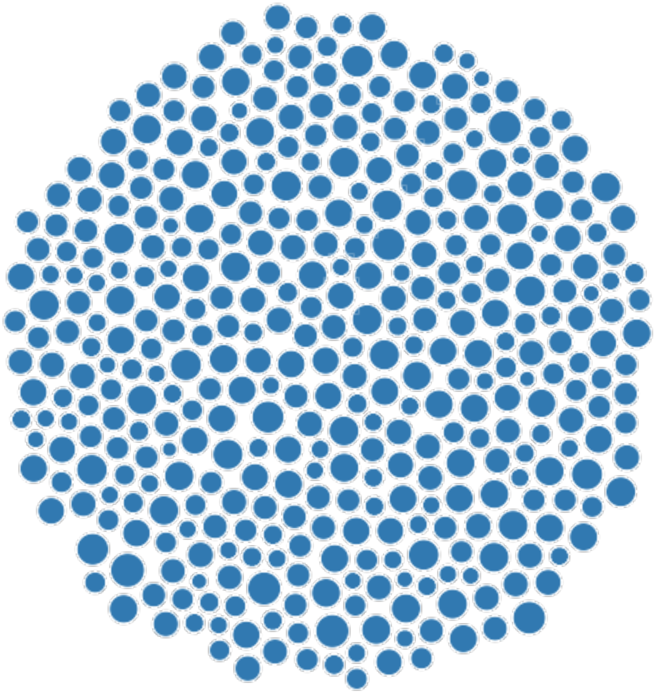


Box Plot

Raw (with Layout Algorithm)

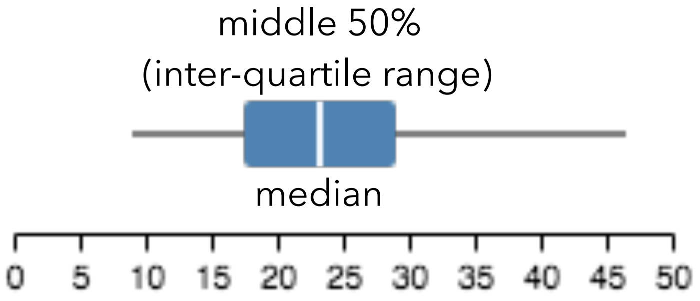


Treemap



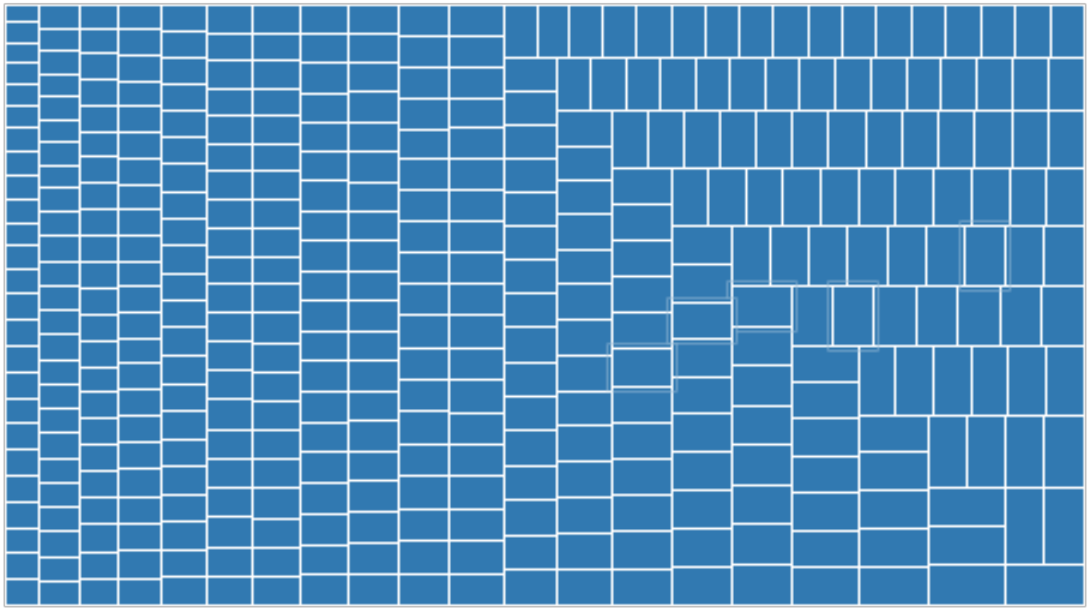
Bubble Chart

Aggregate (Distributions)

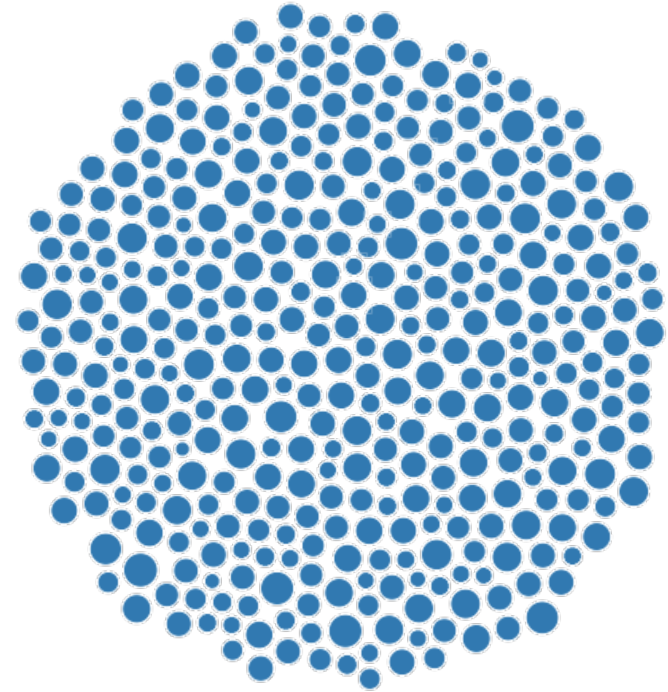


Box Plot

Raw (with Layout Algorithm)

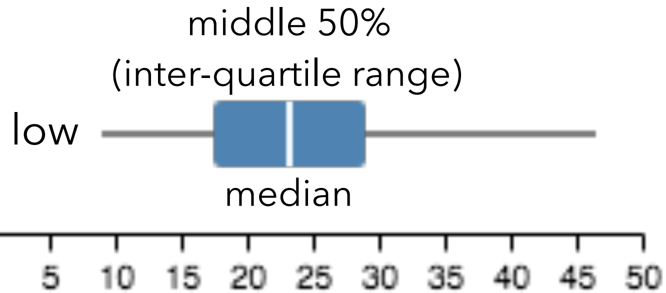


Treemap



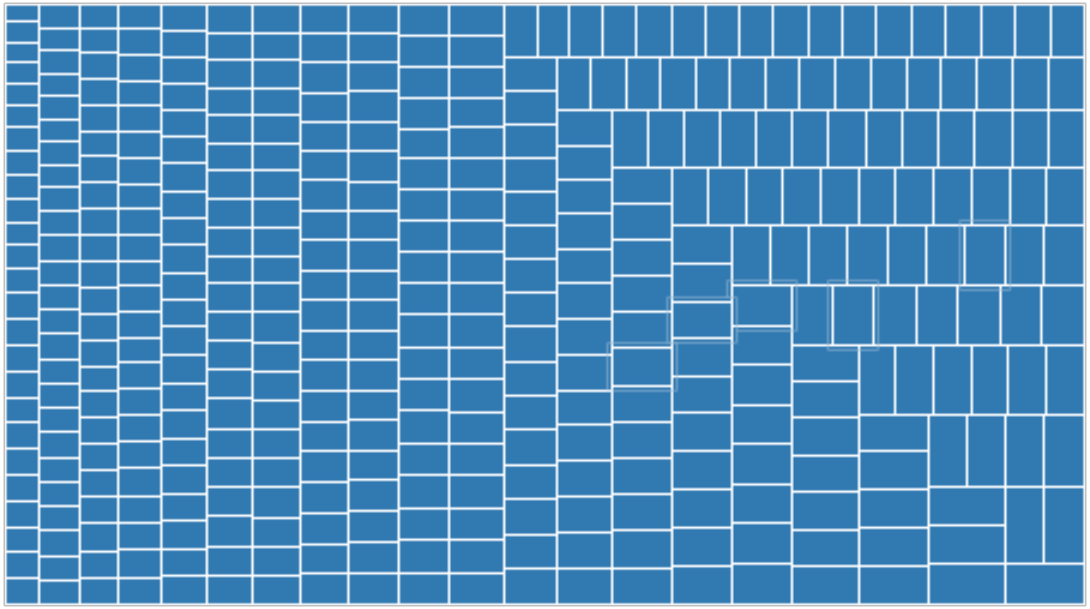
Bubble Chart

Aggregate (Distributions)

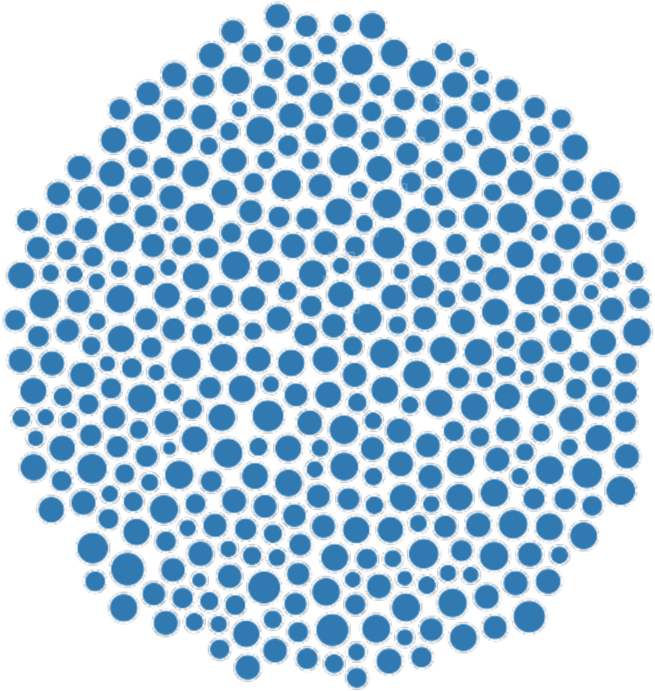


Box Plot

Raw (with Layout Algorithm)

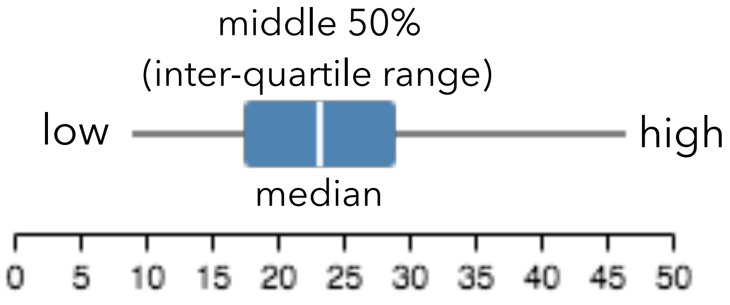


Treemap



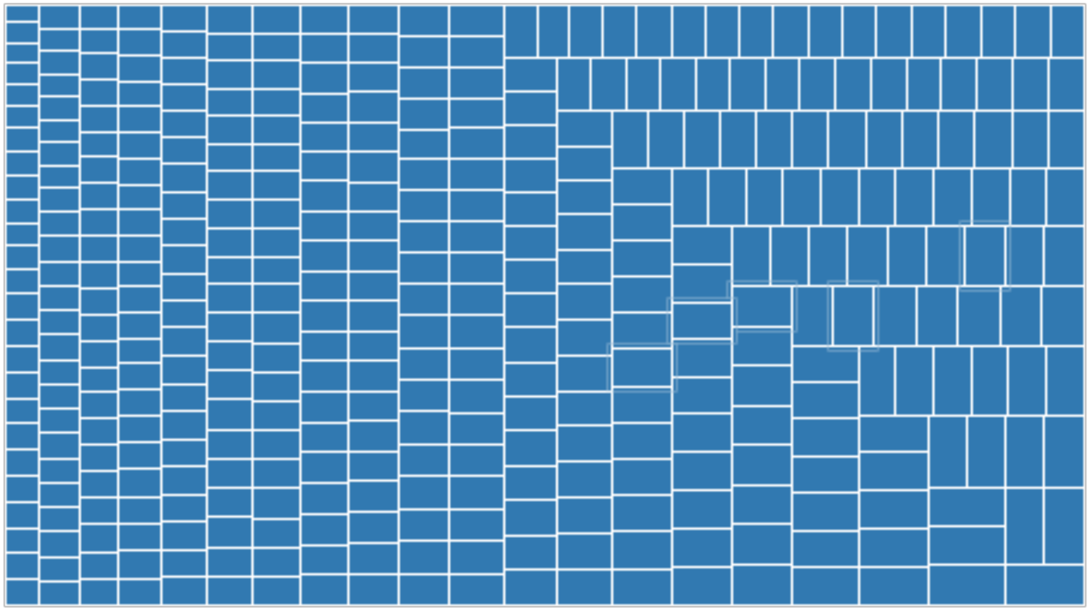
Bubble Chart

Aggregate (Distributions)

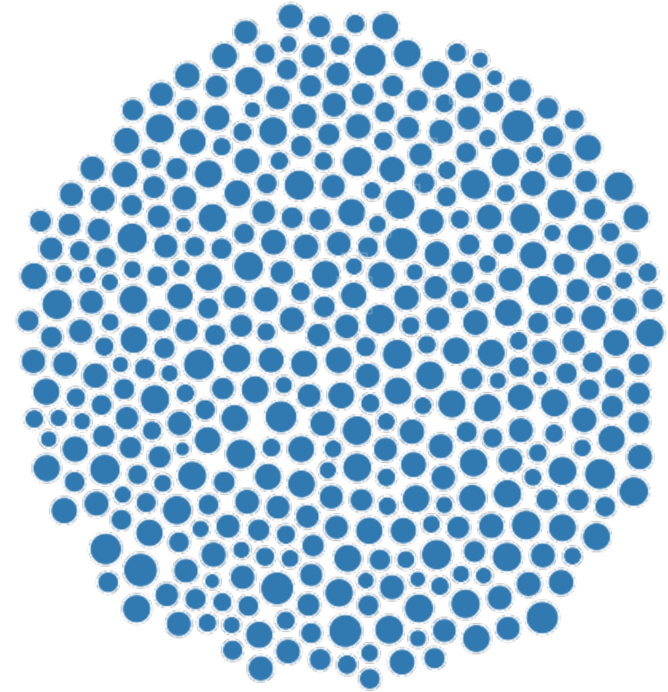


Box Plot

Raw (with Layout Algorithm)

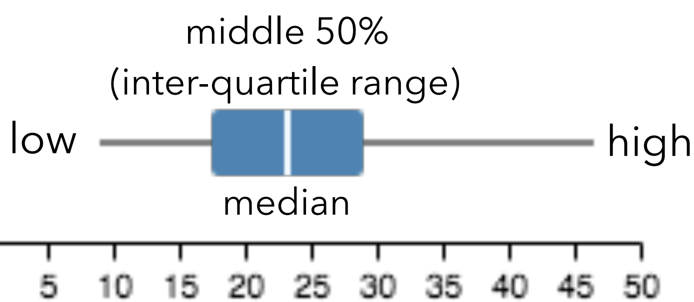


Treemap

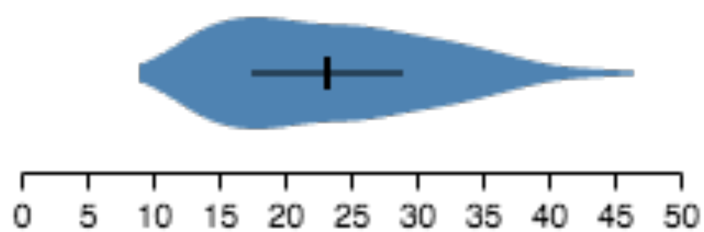


Bubble Chart

Aggregate (Distributions)



Box Plot



Violin Plot

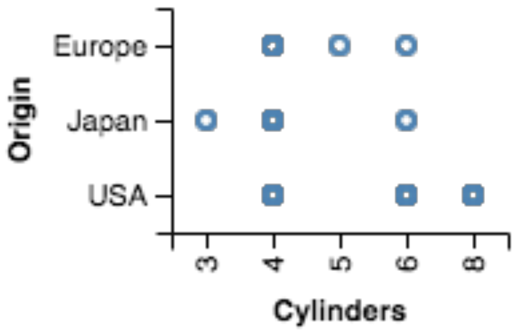
2D: Nominal x Nominal

2D: Nominal x Nominal

Raw

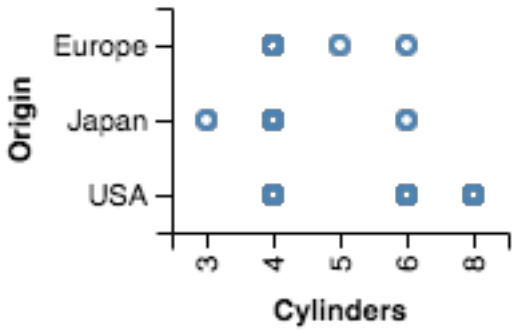
2D: Nominal x Nominal

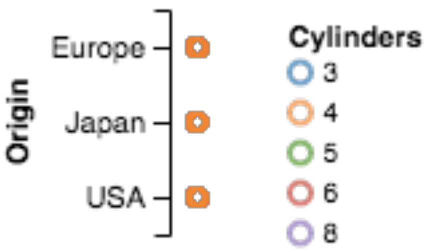
Raw



2D: Nominal x Nominal

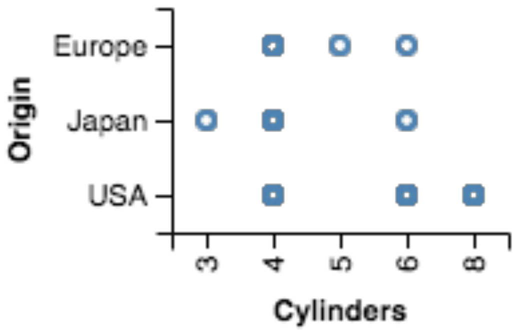
Raw

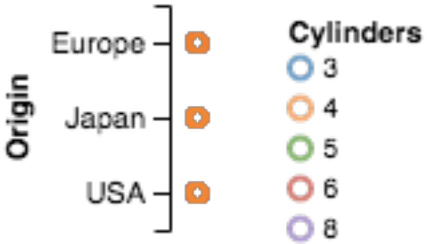


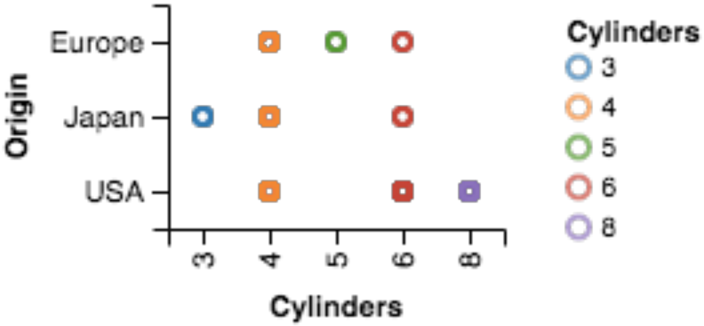


2D: Nominal x Nominal

Raw



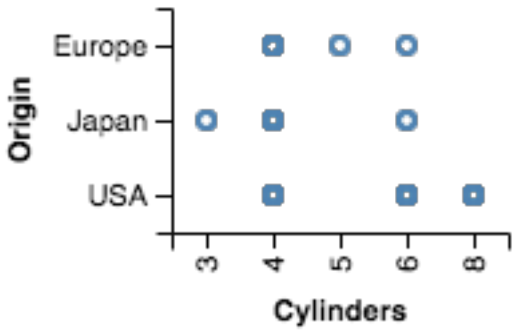


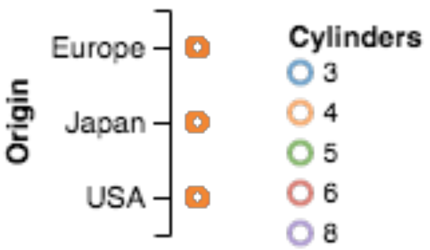


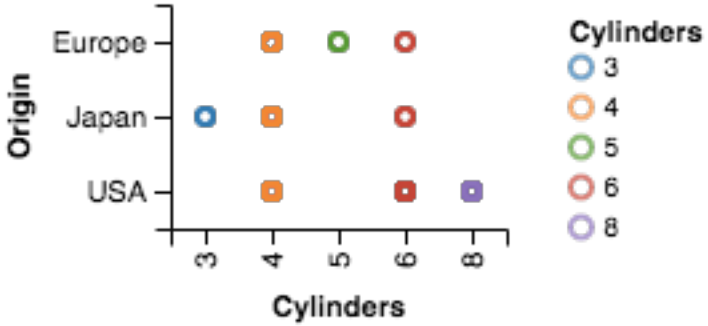
2D: Nominal x Nominal

Raw

Aggregate (Count)

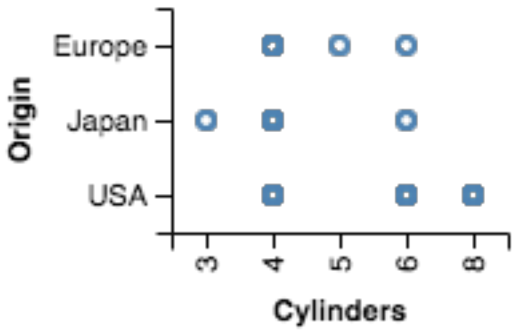


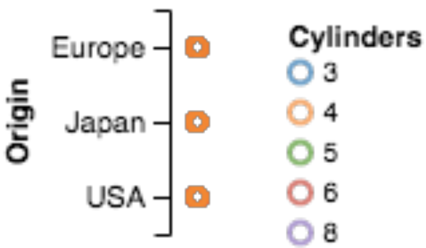


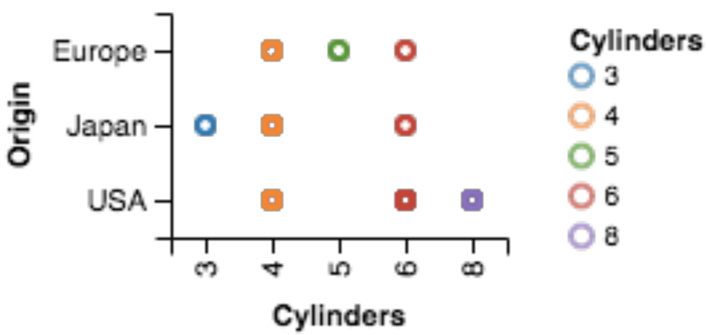


2D: Nominal x Nominal

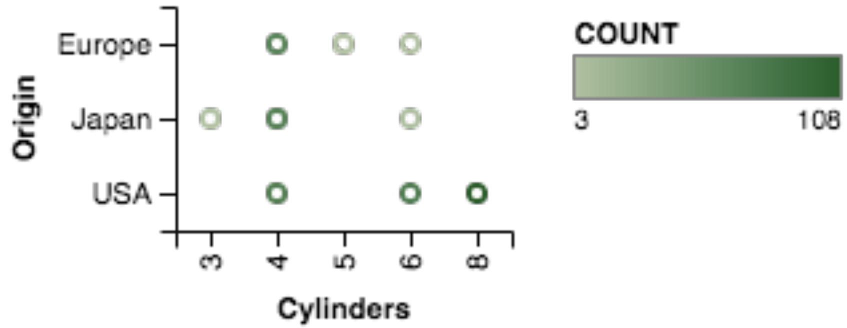
Raw





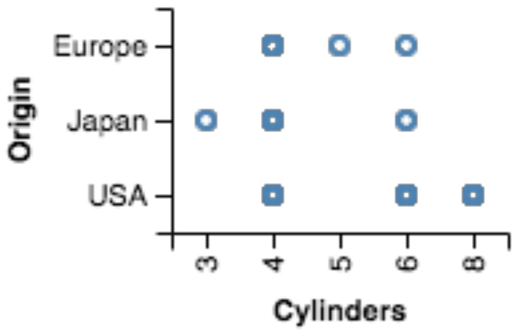


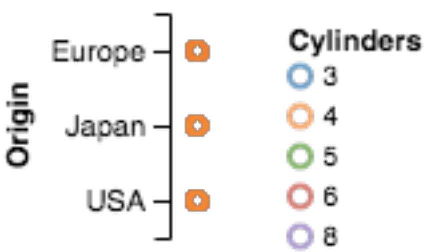
Aggregate (Count)

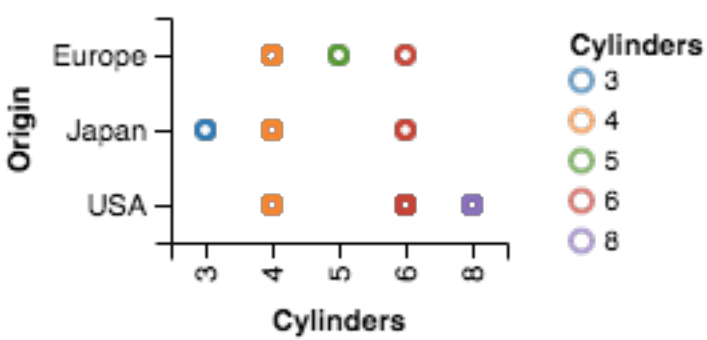


2D: Nominal x Nominal

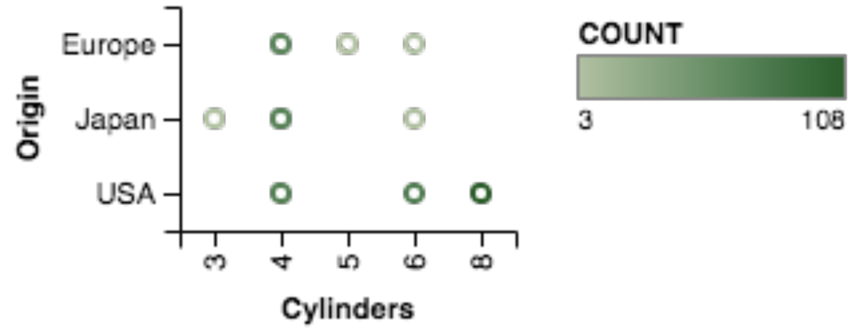
Raw

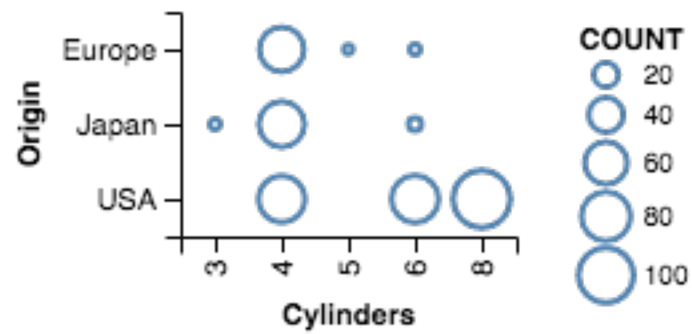






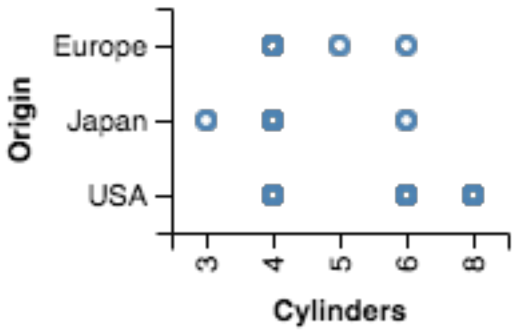
Aggregate (Count)

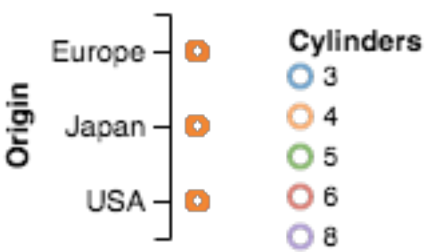


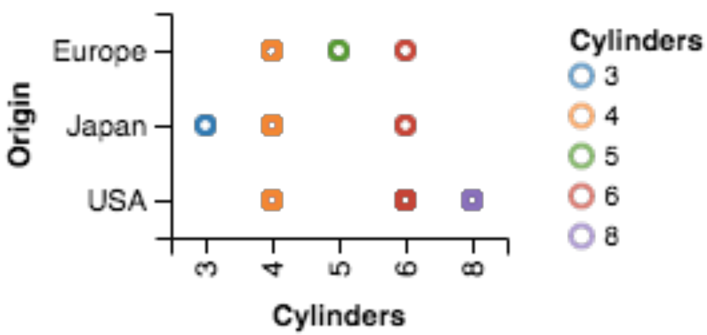


2D: Nominal x Nominal

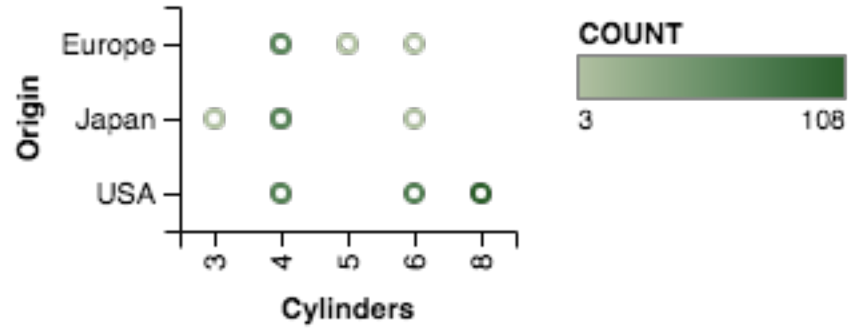
Raw

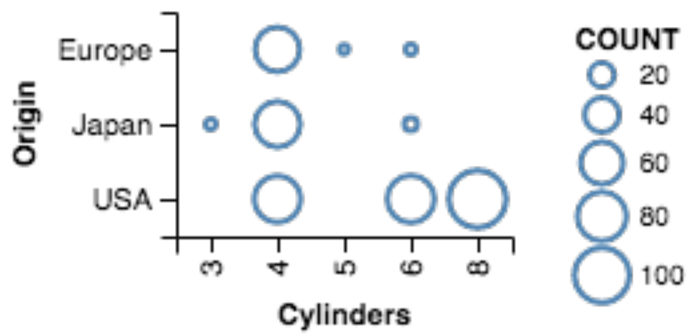


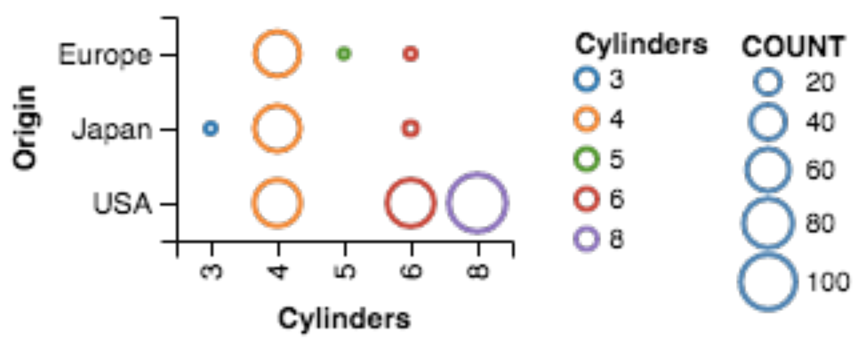




Aggregate (Count)







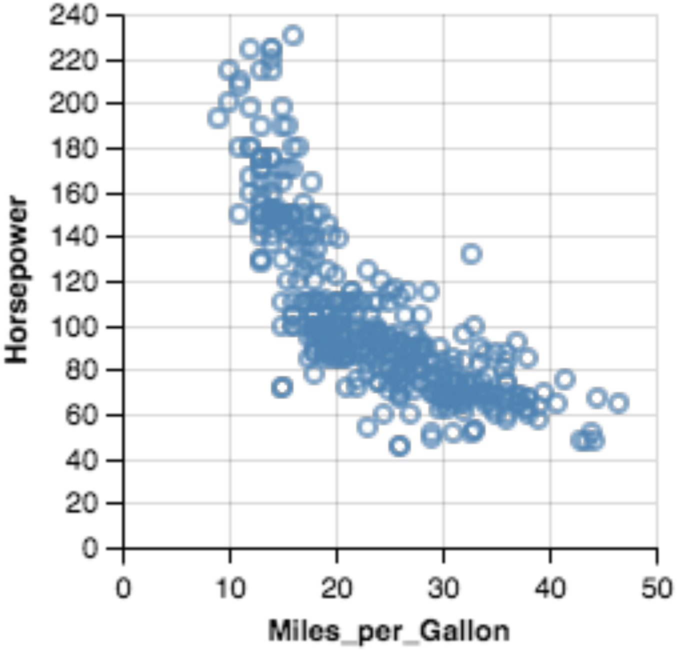
2D: Quantitative x Quantitative

2D: Quantitative x Quantitative

Raw

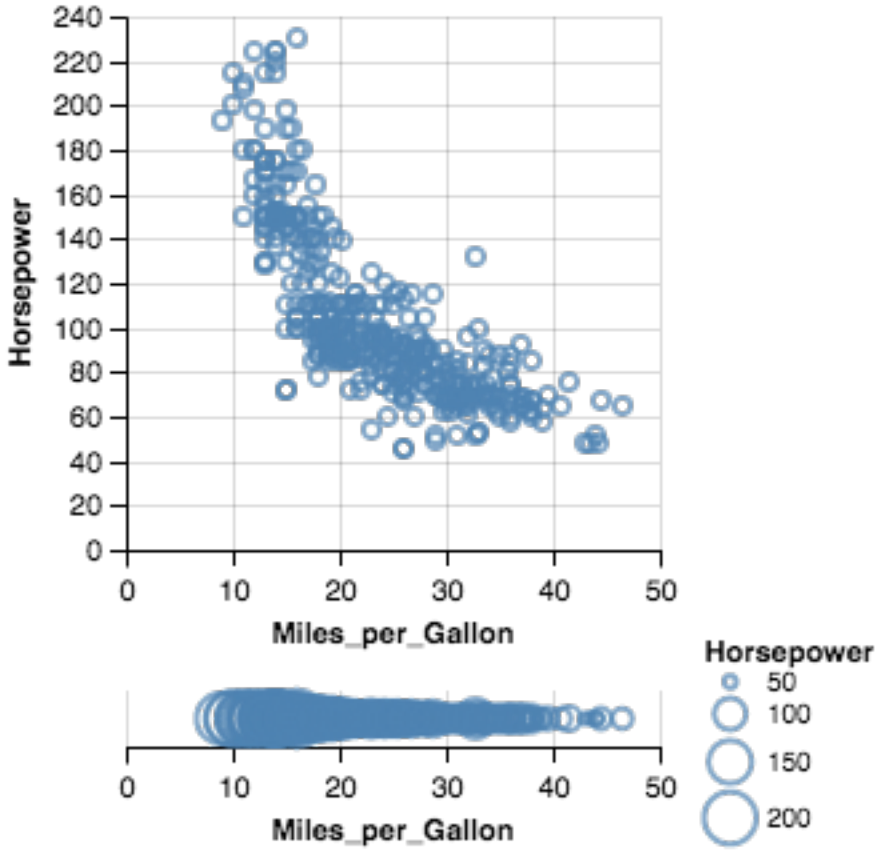
2D: Quantitative x Quantitative

Raw



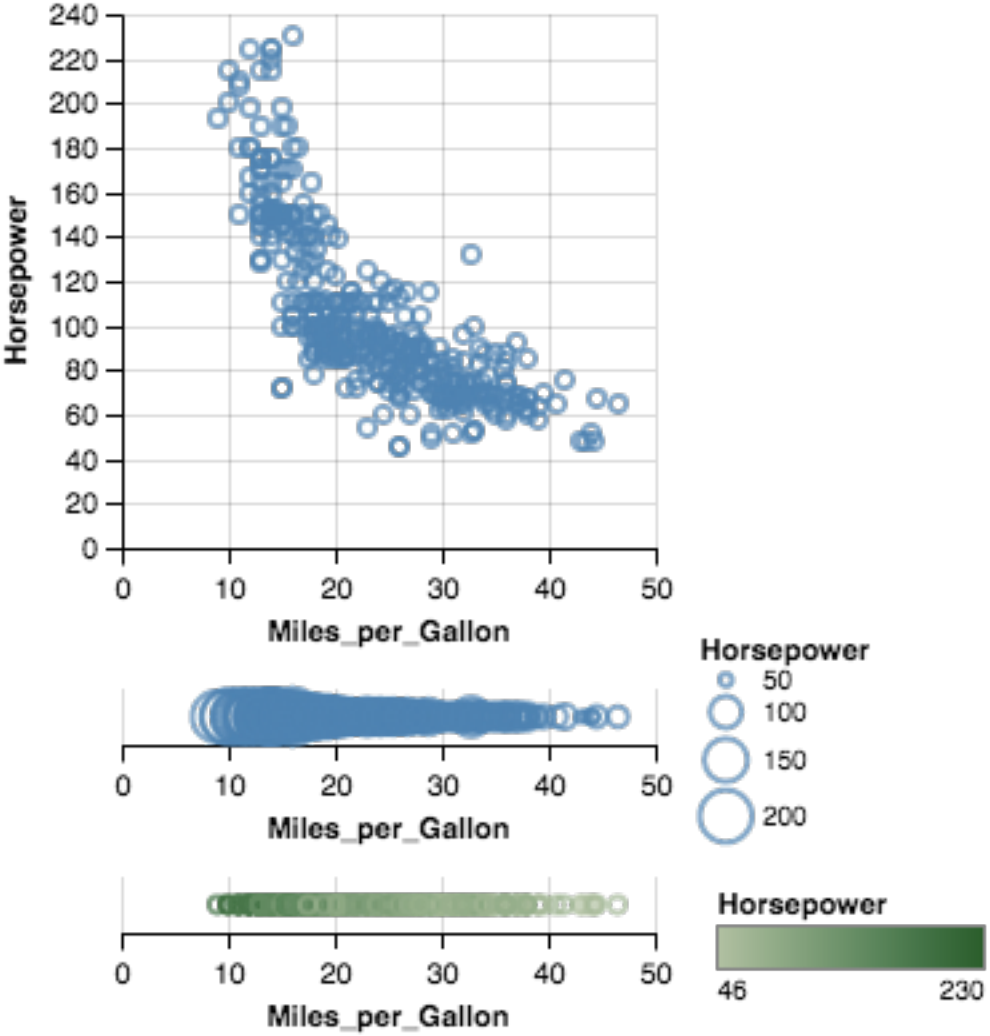
2D: Quantitative x Quantitative

Raw



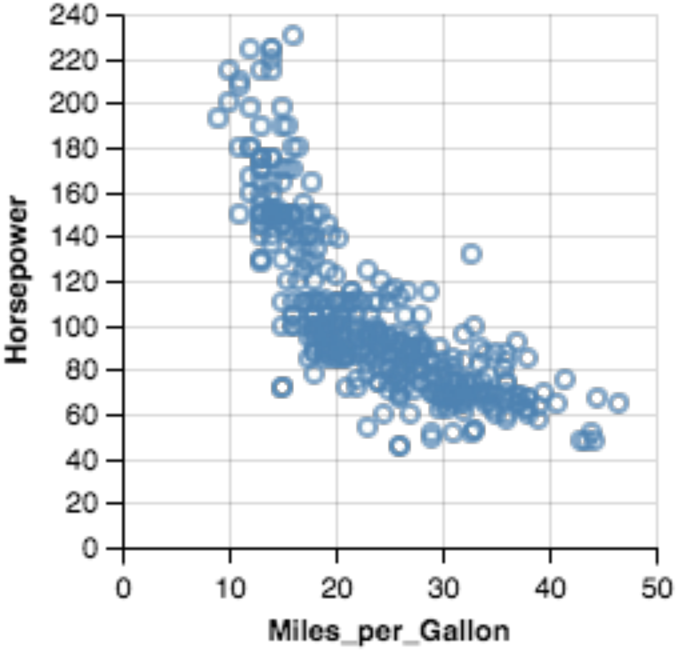
2D: Quantitative x Quantitative

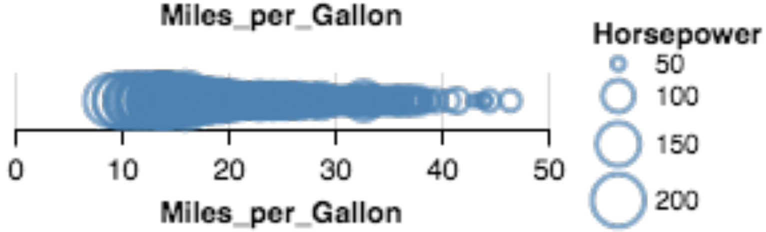
Raw

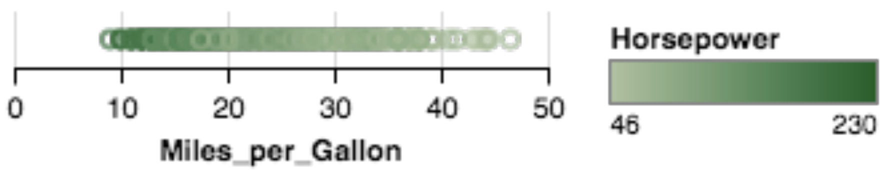


2D: Quantitative x Quantitative

Raw



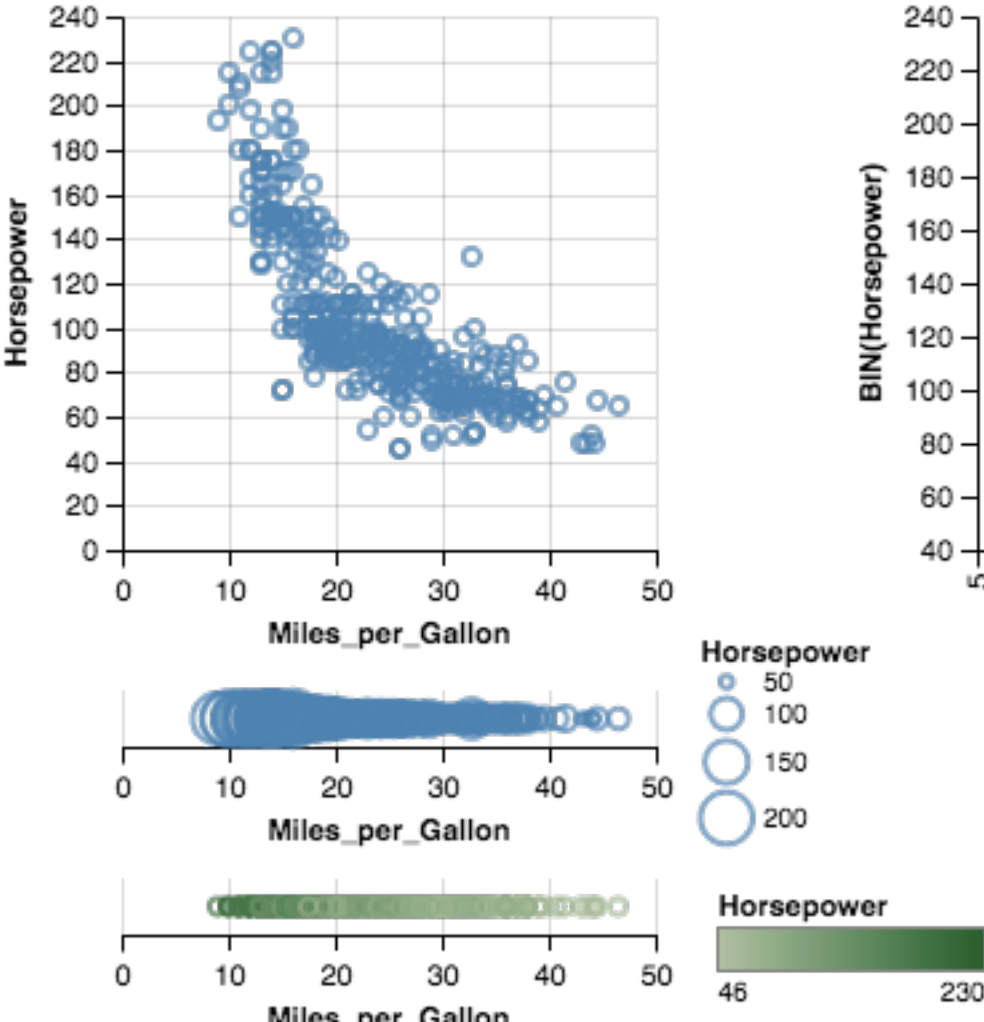




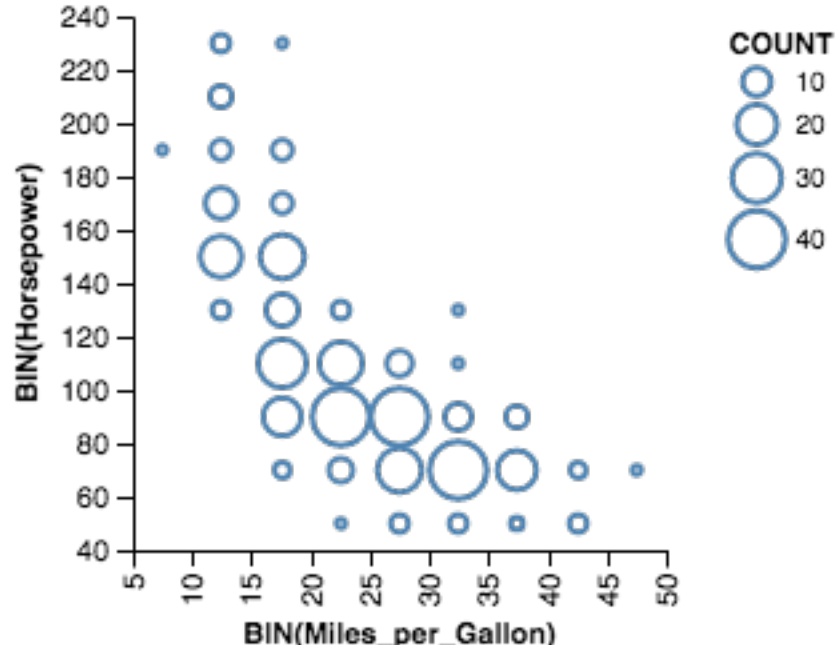
Aggregate (Count)

2D: Quantitative x Quantitative

Raw



Aggregate (Count)



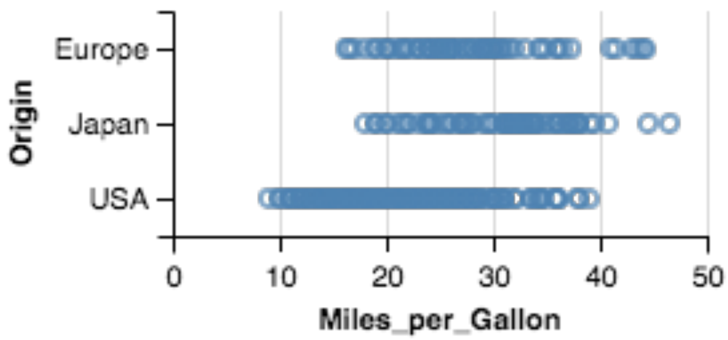
2D: Nominal x Quantitative

2D: Nominal x Quantitative

Raw

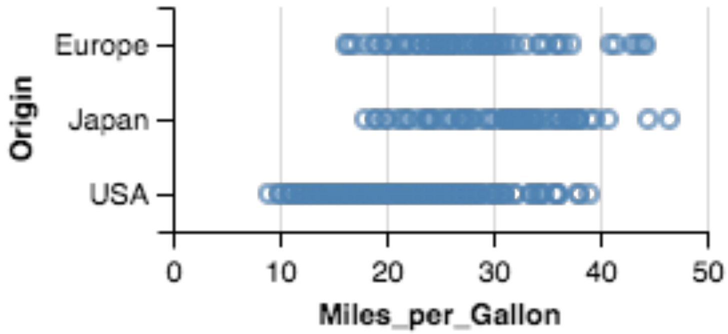
2D: Nominal x Quantitative

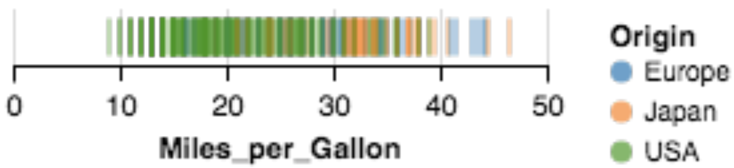
Raw



2D: Nominal x Quantitative

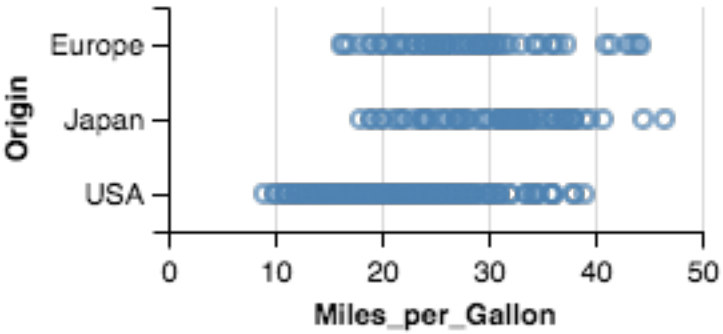
Raw

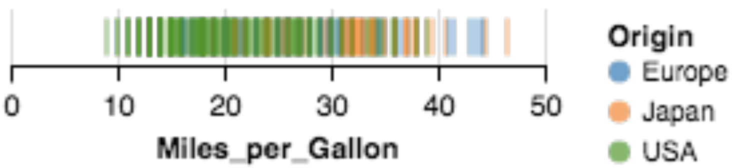


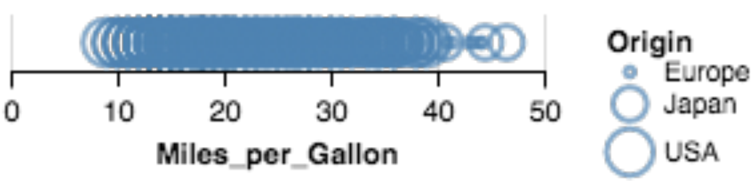


2D: Nominal x Quantitative

Raw

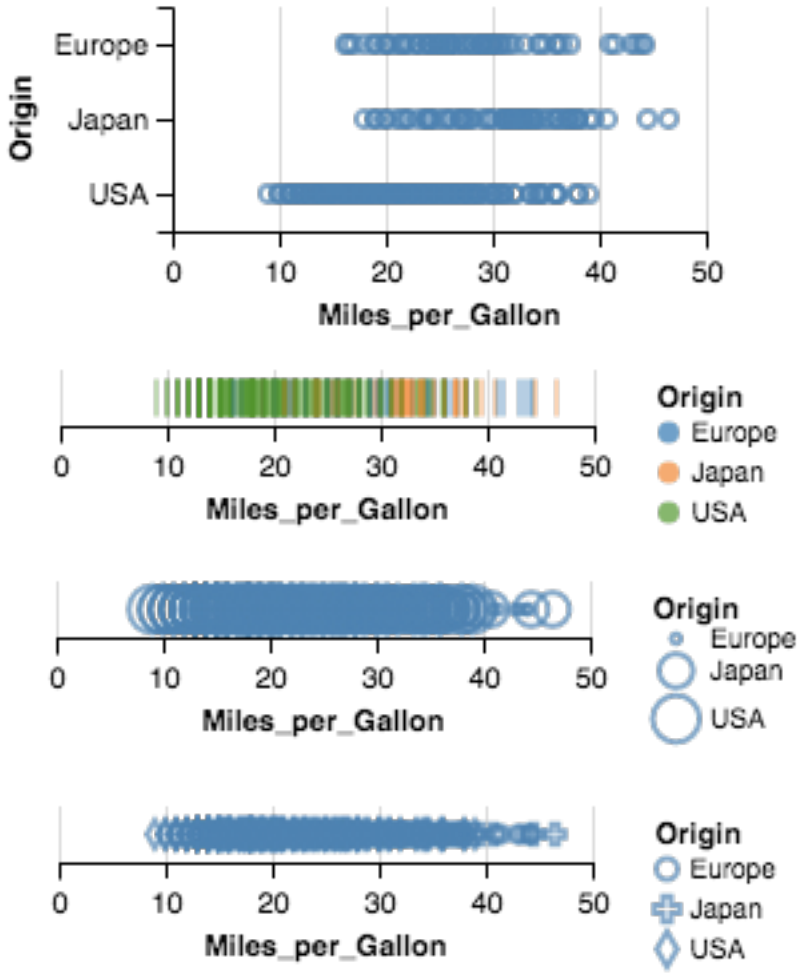






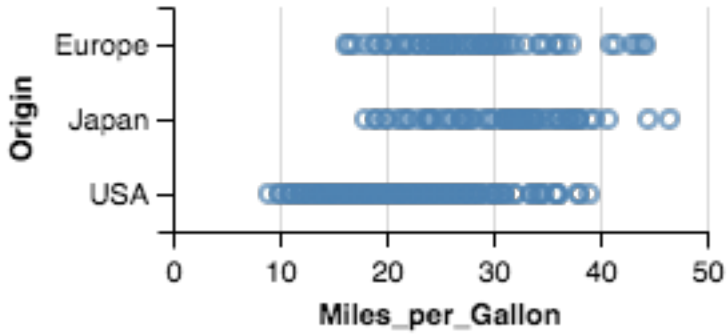
2D: Nominal x Quantitative

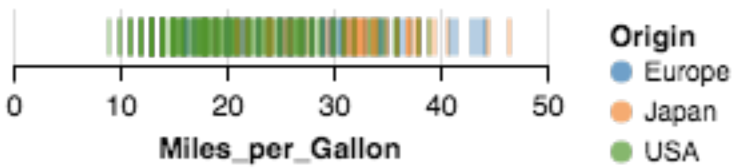
Raw

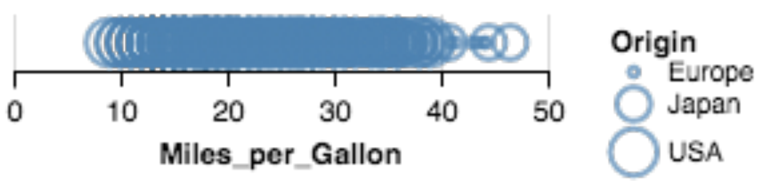


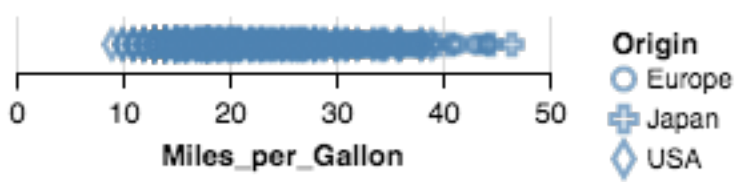
2D: Nominal x Quantitative

Raw





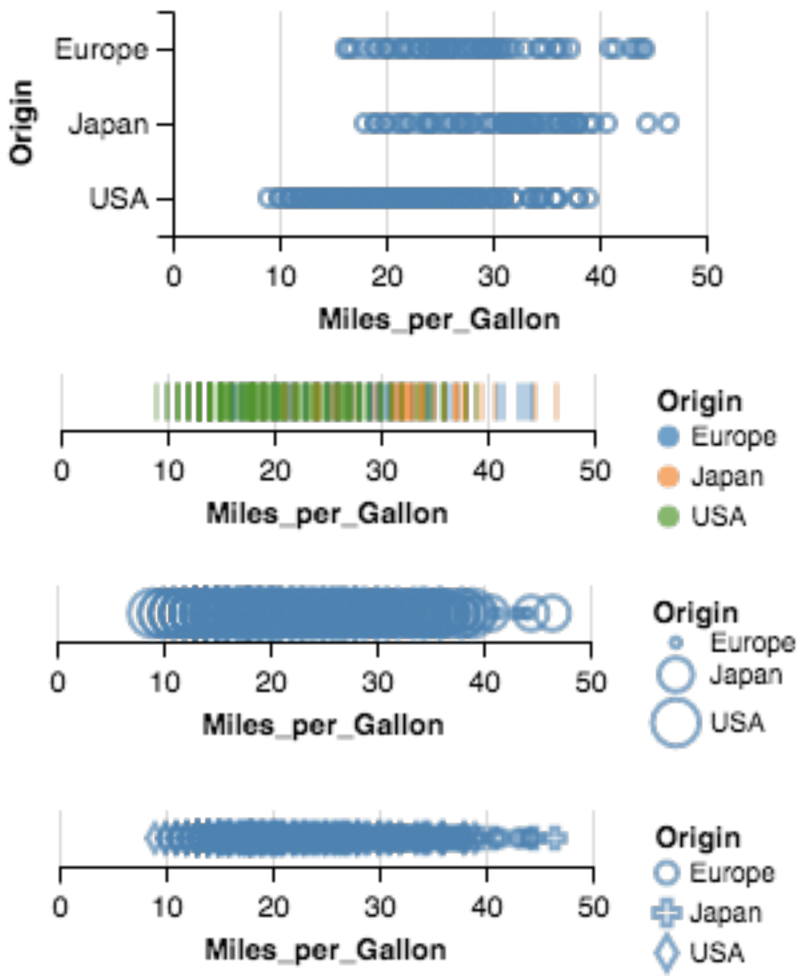




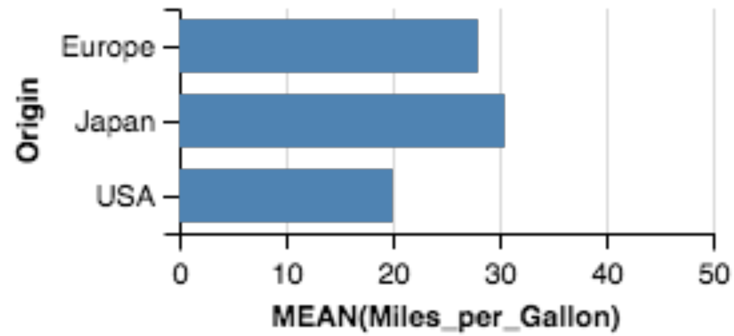
Aggregate (Mean)

2D: Nominal x Quantitative

Raw

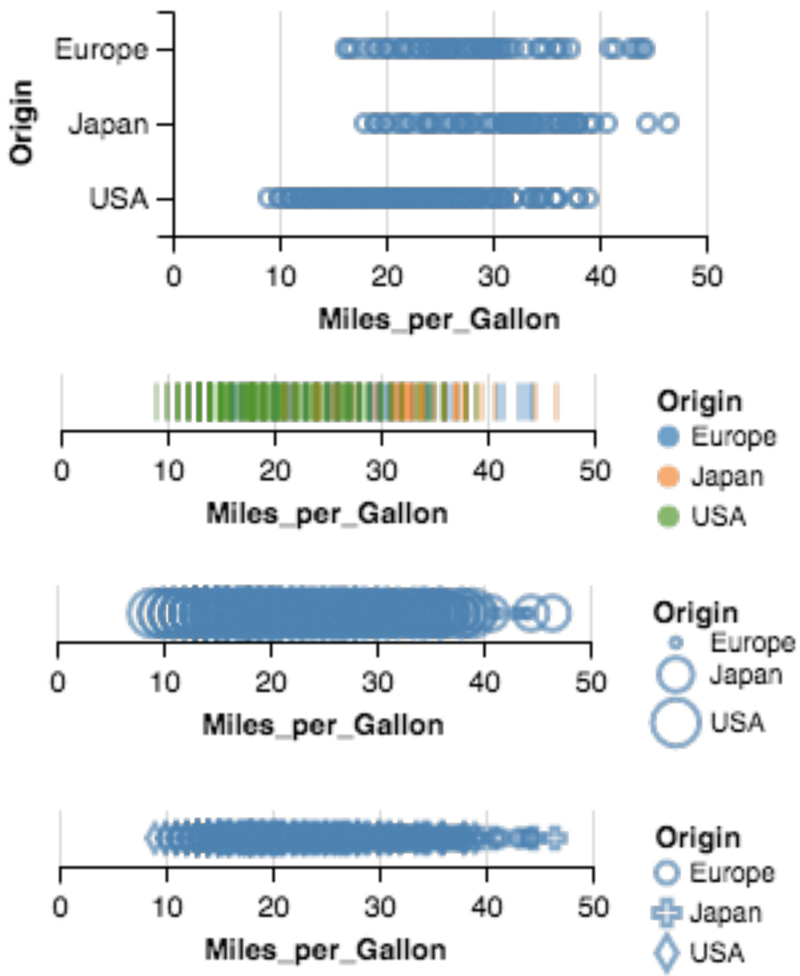


Aggregate (Mean)

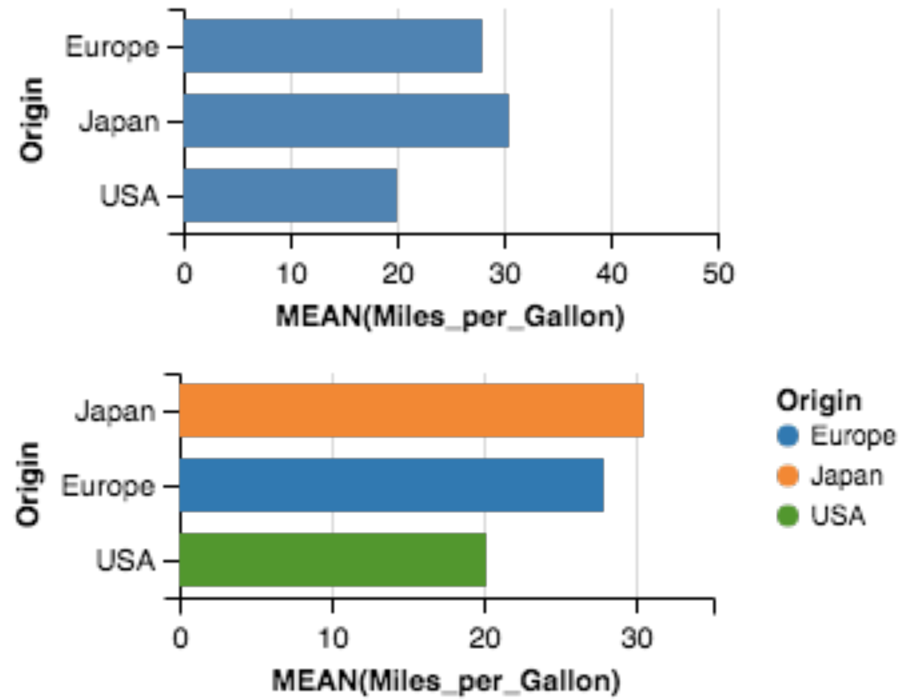


2D: Nominal x Quantitative

Raw

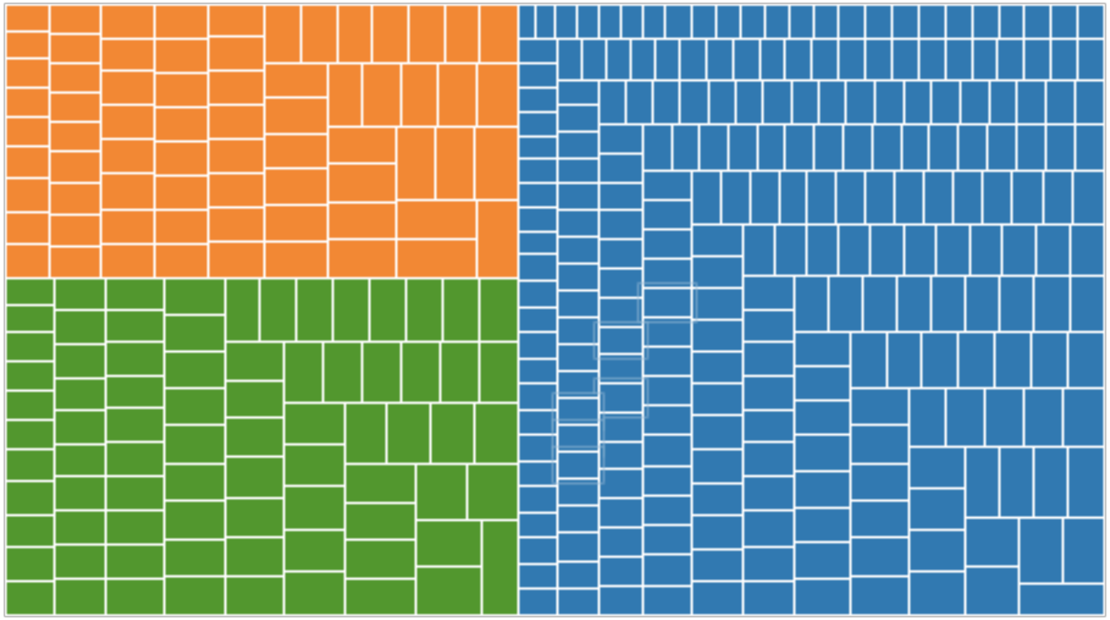


Aggregate (Mean)



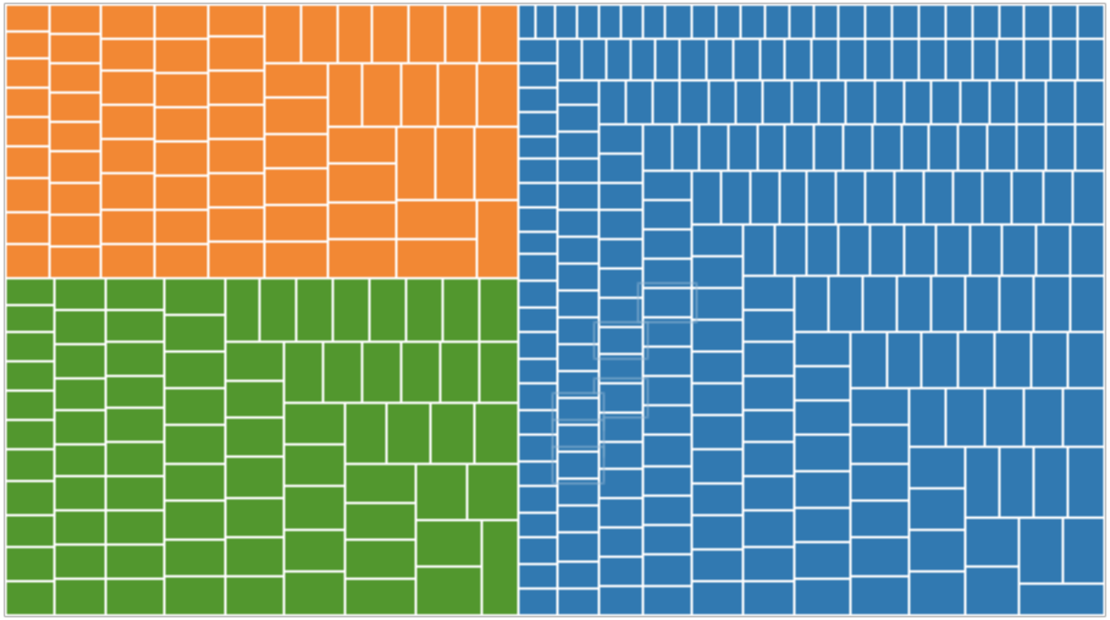
Raw (with Layout Algorithm)

Raw (with Layout Algorithm)

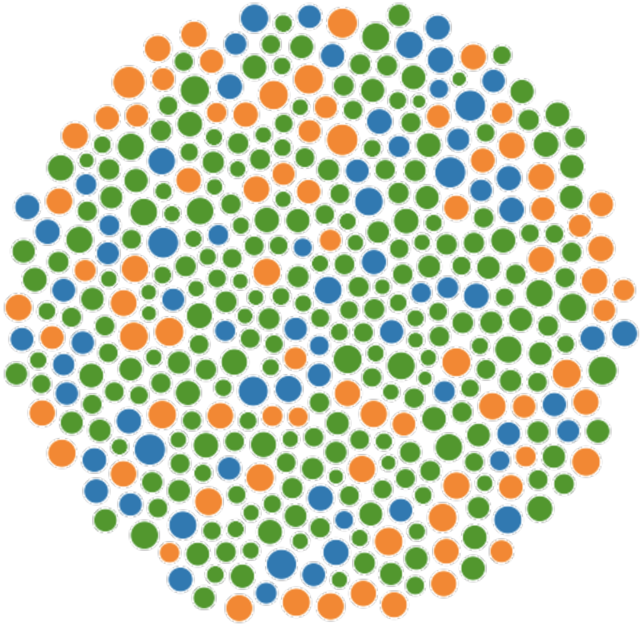


Treemap

Raw (with Layout Algorithm)

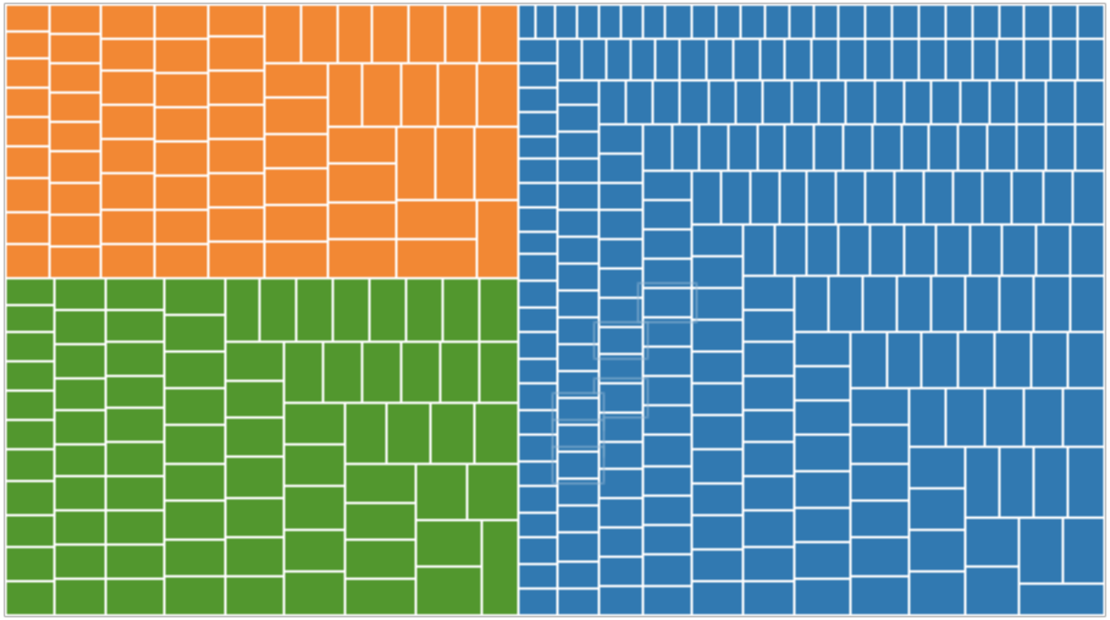


Treemap

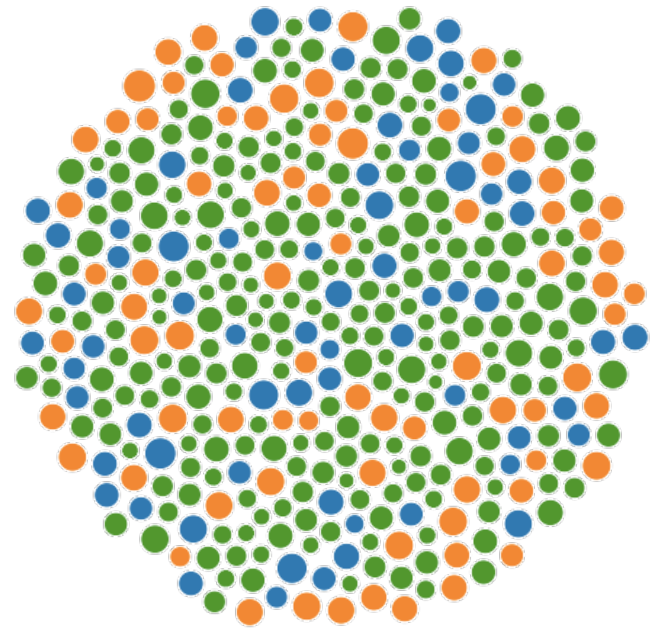


Bubble Chart

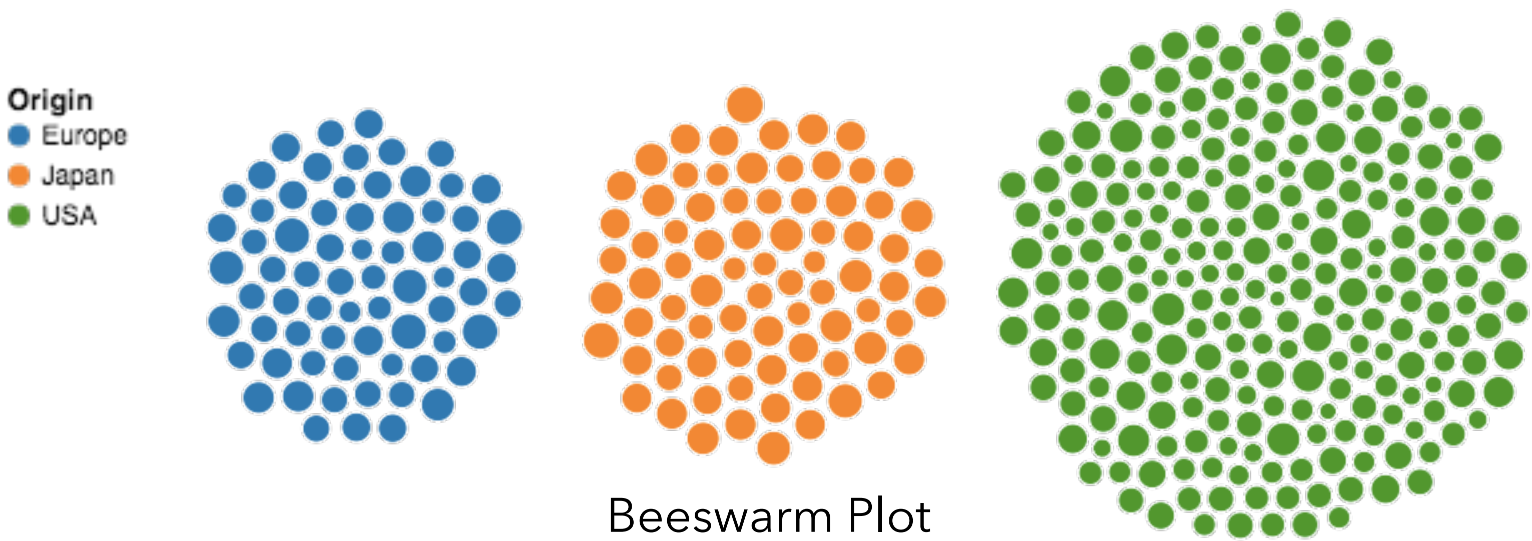
Raw (with Layout Algorithm)



Treemap



Bubble Chart



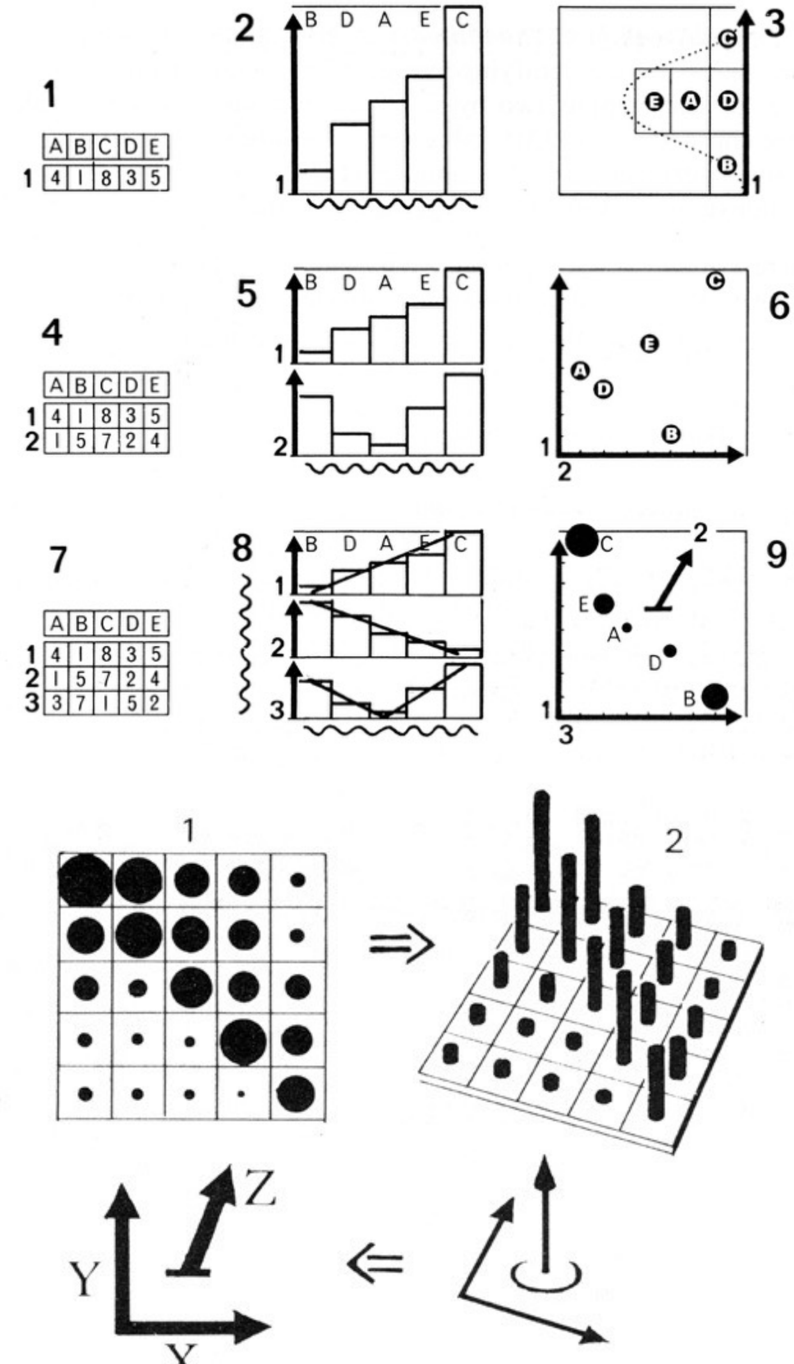
Beeswarm Plot

3D and Higher

Two variables [x,y]
Can map to 2D points.
Scatterplots, maps, …

Third variable [z]
Often use one of size, color, opacity, shape, *etc*. Or, one can further partition space.

What about 3D rendering?



[Bertin]

Other Visual Encoding Channels?

Encoding Effectiveness

Effectiveness Rankings [Mackinlay 86]

QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density (Value)	Color Hue
Angle	Color Sat	Texture
Slope	Color Hue	Connection
Area (Size)	Texture	Containment
Volume	Connection	Density (Value)
Density (Value)	Containment	Color Sat
Color Sat	Length	Shape
Color Hue	Angle	Length
Texture	Slope	Angle
Connection	Area (Size)	Slope
Containment	Volume	Area
Shape	Shape	Volume

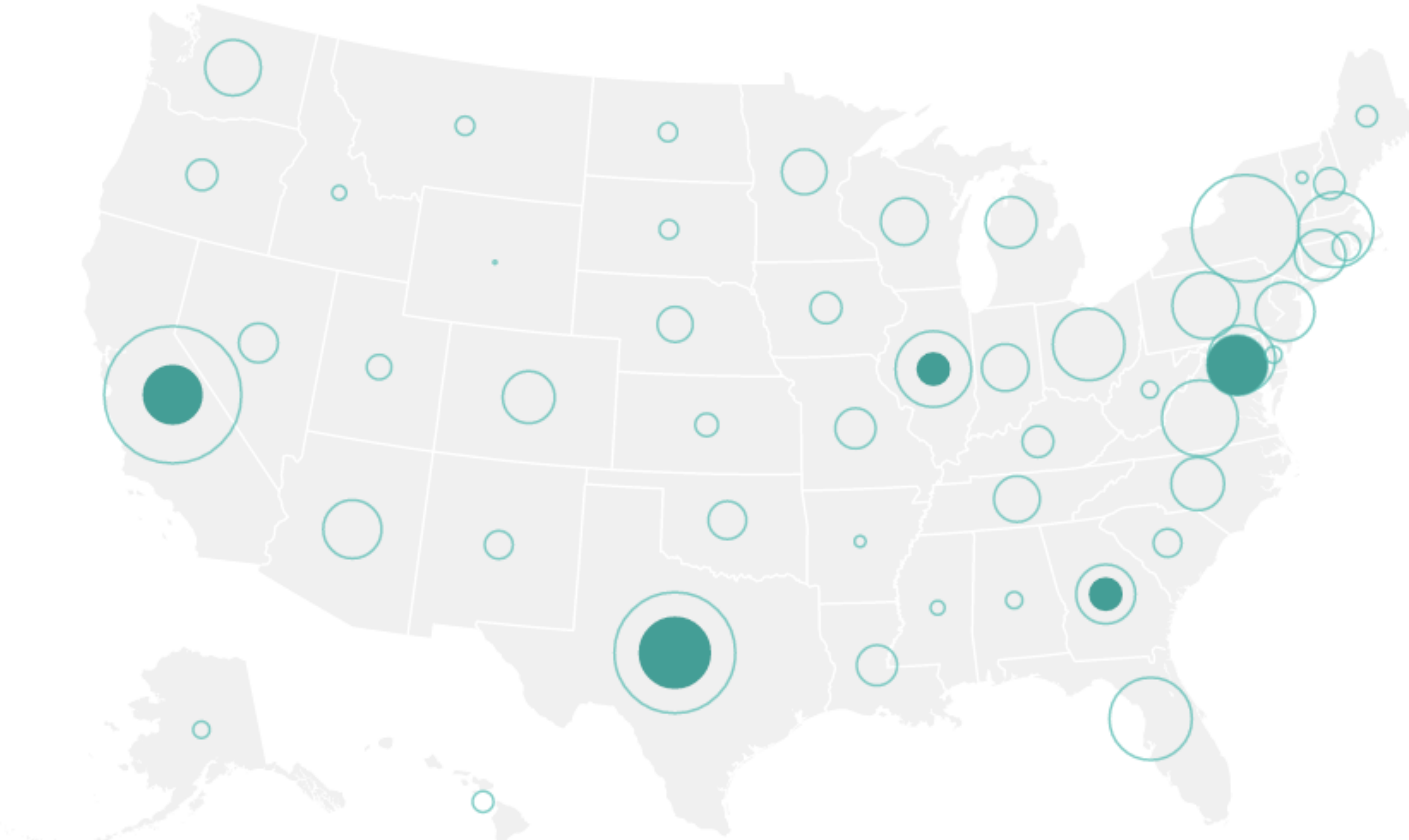
Effectiveness Rankings [Mackinlay 86]

QUANTITATIVE	ORDINAL	NOMINAL
Position ············	**Position** ············	**Position**
Length	Density (Value)	Color Hue
Angle	Color Sat	Texture
Slope	Color Hue	Connection
Area (Size)	Texture	Containment
Volume	Connection	Density (Value)
Density (Value)	Containment	Color Sat
Color Sat	Length	Shape
Color Hue	Angle	Length
Texture	Slope	Angle
Connection	Area (Size)	Slope
Containment	Volume	Area
Shape	Shape	Volume

Effectiveness Rankings [Mackinlay 86]

QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density (Value)	**Color Hue**
Angle	Color Sat	Texture
Slope	**Color Hue**	Connection
Area (Size)	Texture	Containment
Volume	Connection	Density (Value)
Density (Value)	Containment	Color Sat
Color Sat	Length	Shape
Color Hue	Angle	Length
Texture	Slope	Angle
Connection	Area (Size)	Slope
Containment	Volume	Area
Shape	Shape	Volume

Color Encoding



Area Encoding

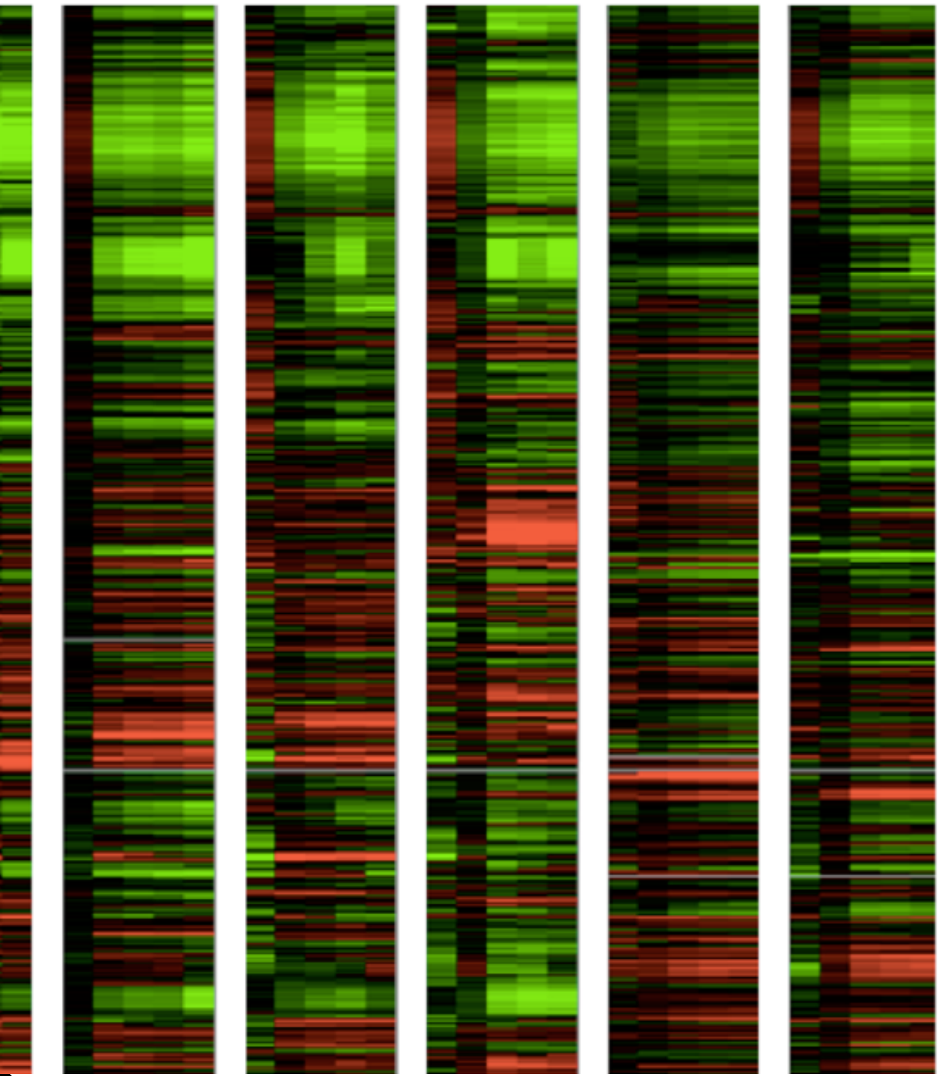
Effectiveness Rankings

QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density (Value)	Color Hue
Angle	Color Sat	Texture
Slope	Color Hue	Connection
Area (Size)	Texture	Containment
Volume	Connection	Density (Value)
~~**Density (Value)**~~	Containment	Color Sat
Color Sat	Length	Shape
Color Hue	Angle	Length
Texture	Slope	Angle
Connection	Area (Size)	Slope
Containment	Volume	Area
Shape	Shape	Volume

Gene Expression Time-Series [Meyer et al '11]

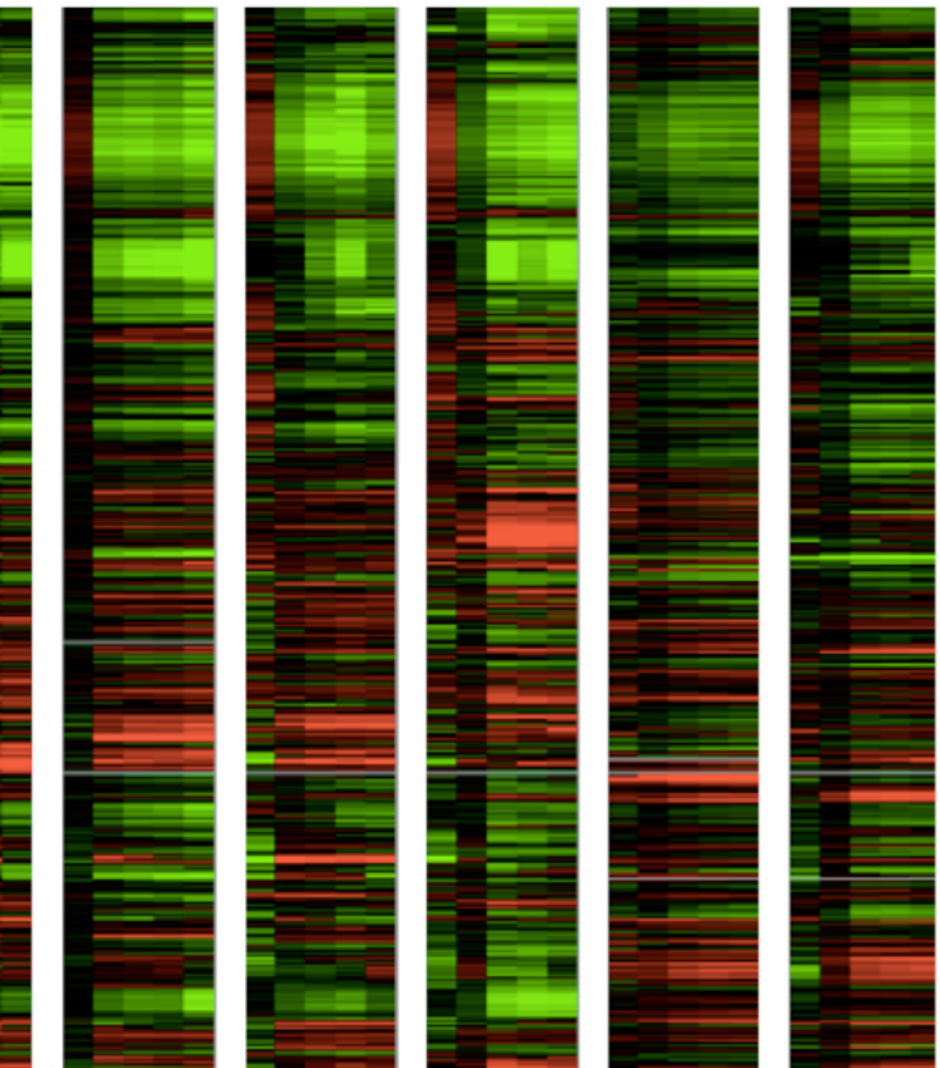
Gene Expression Time-Series [Meyer et al ’11]

Color Encoding

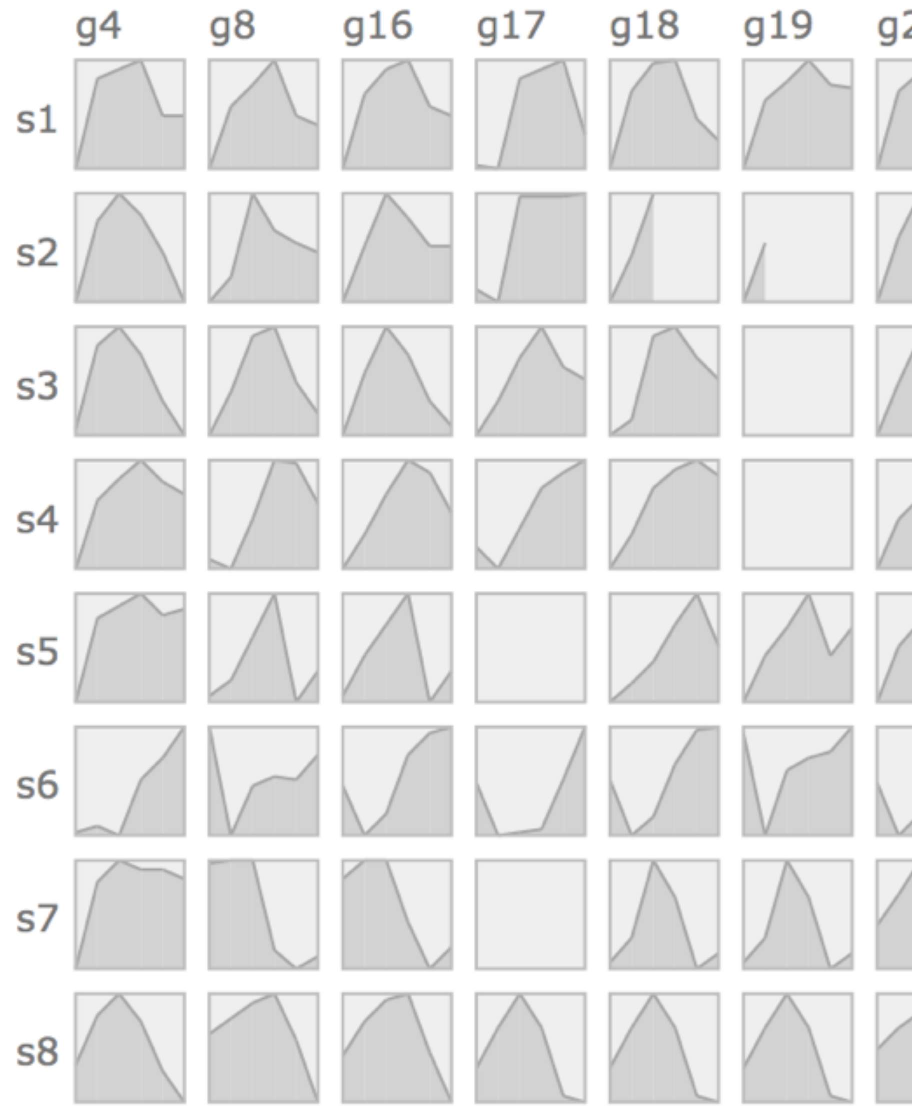


Gene Expression Time-Series [Meyer et al ’11]

Color Encoding



Position Encoding



Effectiveness Rankings

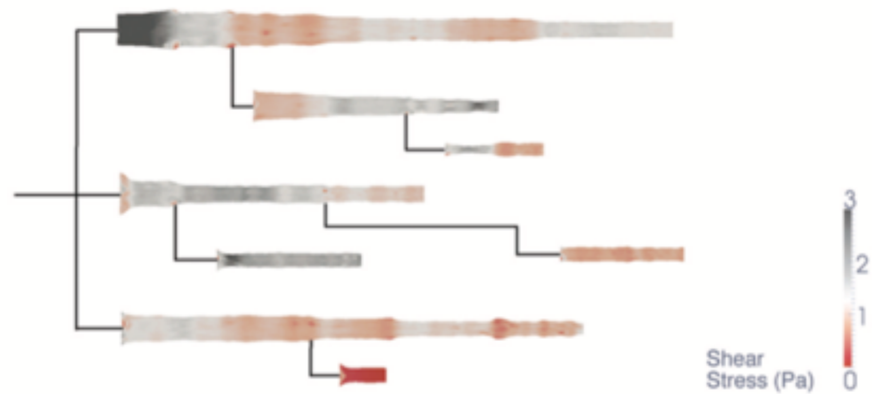
QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density (Value)	Color Hue
Angle	Color Sat	Texture
Slope	Color Hue	Connection
Area (Size)	Texture	Containment
Volume	Connection	Density (Value)
~~**Density (Value)**~~	Containment	Color Sat
Color Sat	Length	Shape
~~**Color Hue**~~	Angle	Length
Texture	Slope	Angle
Connection	Area (Size)	Slope
Containment	Volume	Area
Shape	Shape	Volume

Artery Visualization [Borkin et al '11]

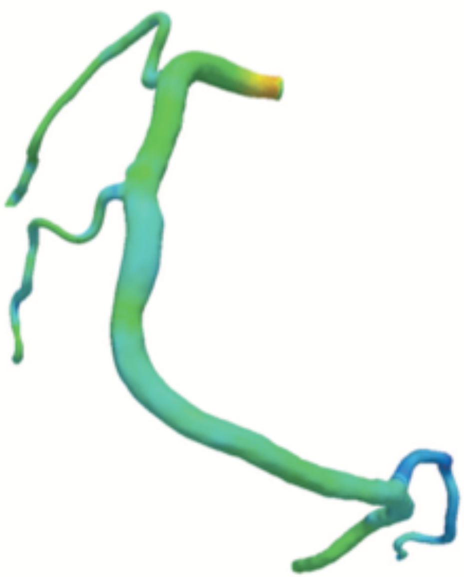
Rainbow Palette

Diverging Palette

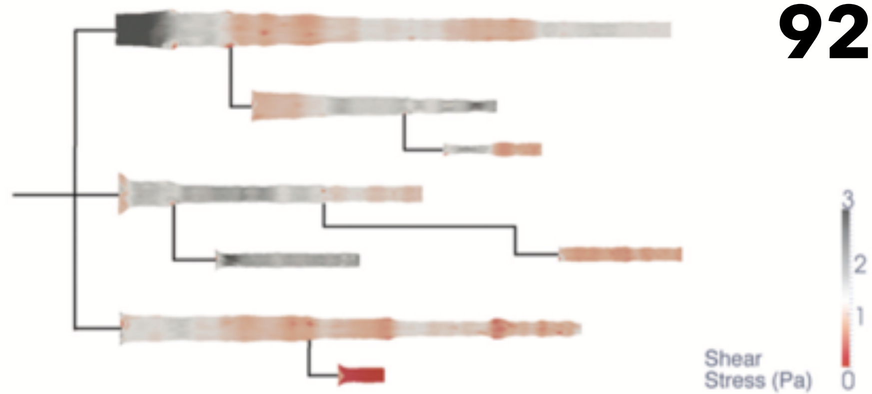
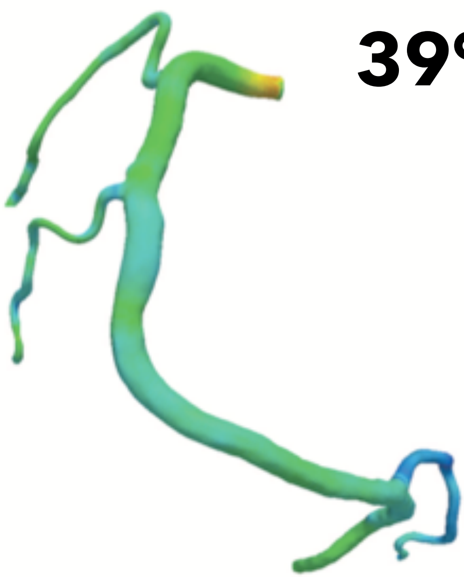
2D



3D



Artery Visualization [Borkin et al '11]

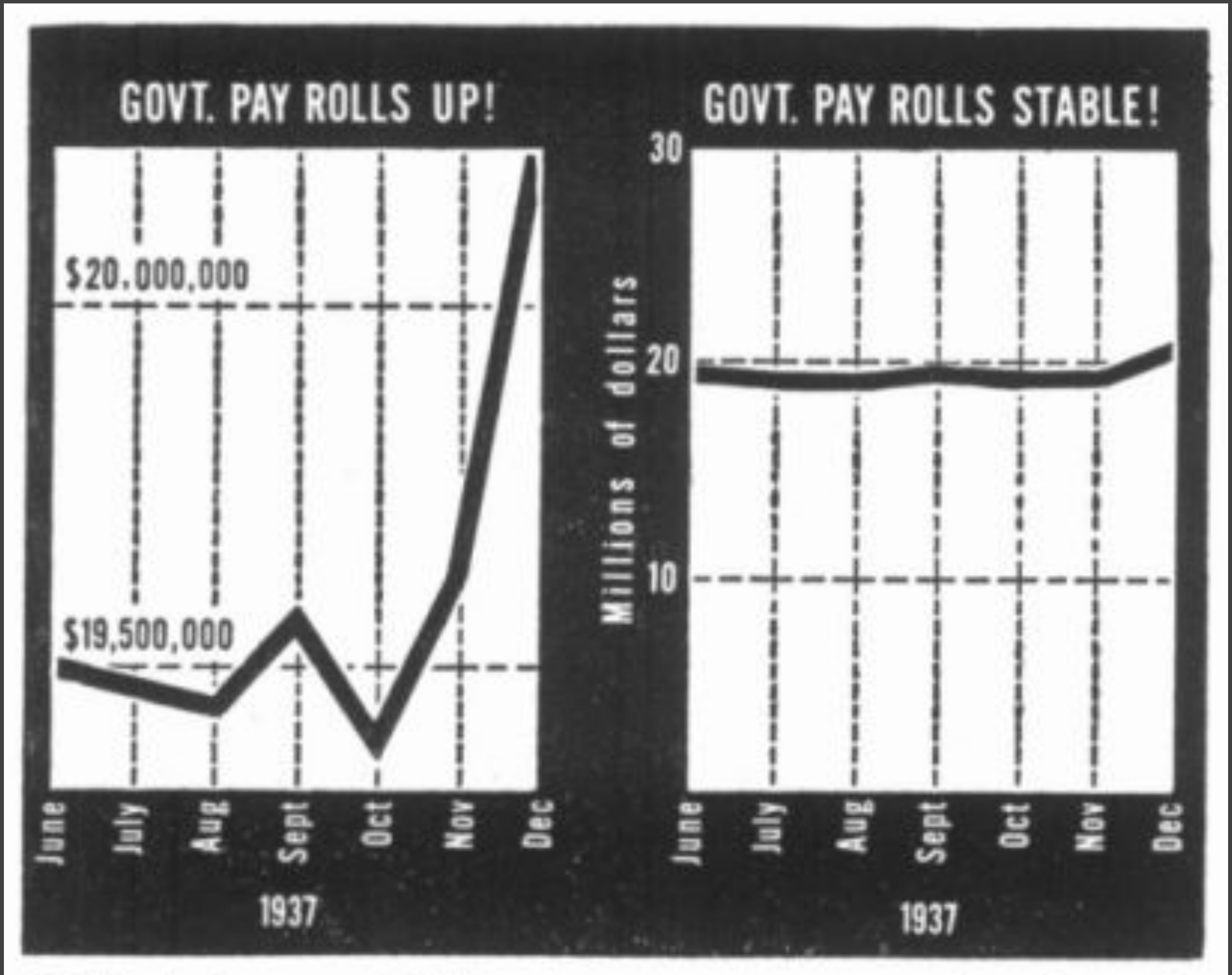
	Rainbow Palette	Diverging Palette
2D	62% 	92% 
3D	39% 	71% 

Effectiveness Rankings

QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density (Value)	Color Hue
Angle	Color Sat	Texture
Slope	Color Hue	Connection
Area (Size)	Texture	Containment
Volume	**Connection**	Density (Value)
Density (Value)	Containment	Color Sat
Color Sat	Length	Shape
~~**Color Hue**~~	Angle	Length
Texture	Slope	Angle
Connection	Area (Size)	Slope
Containment	Volume	Area
Shape	Shape	Volume

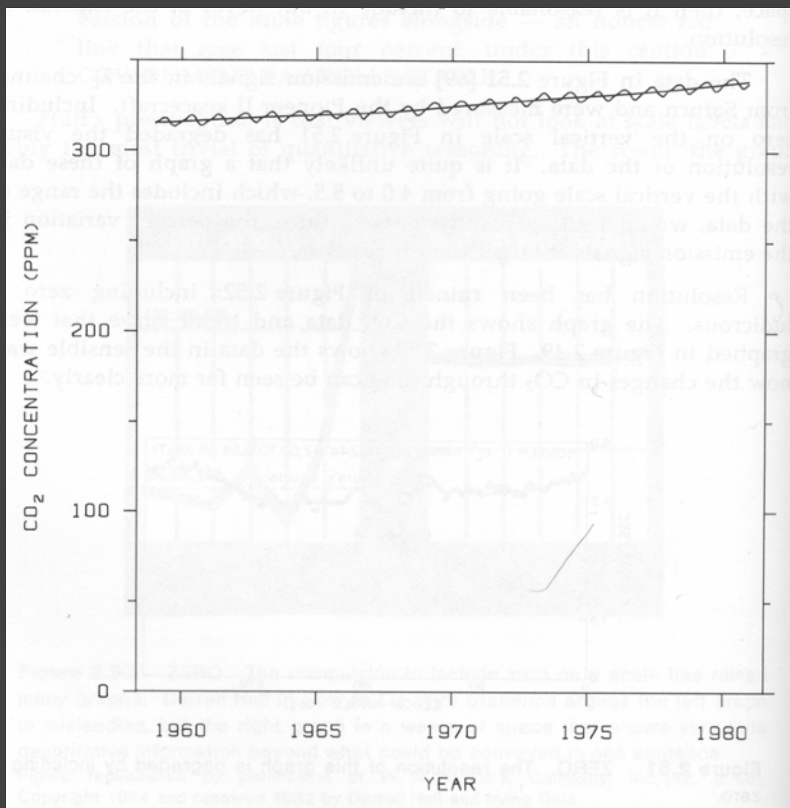
Scales & Axes

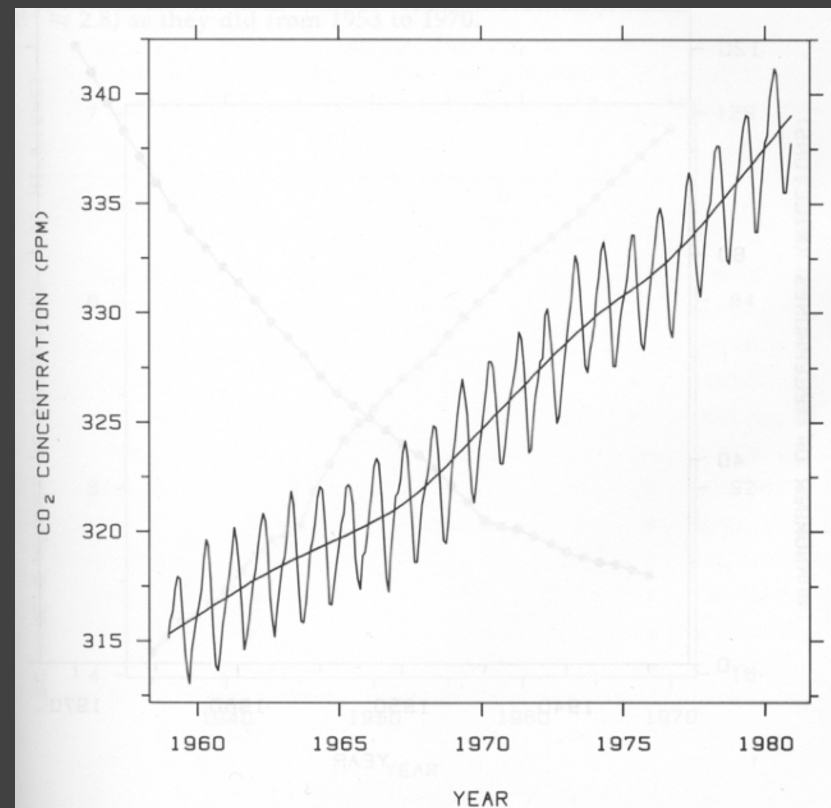
Include Zero in Axis Scale?



Government payrolls in 1937 [How To Lie With Statistics. Huff]

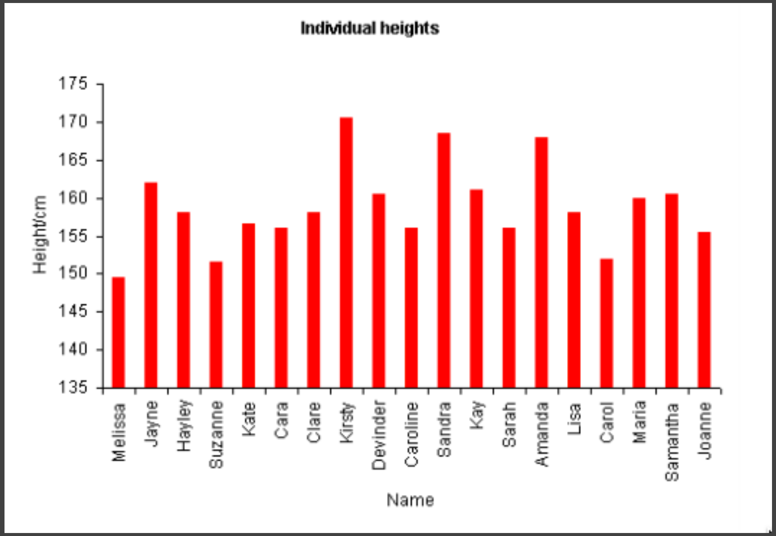
Include Zero in Axis Scale?



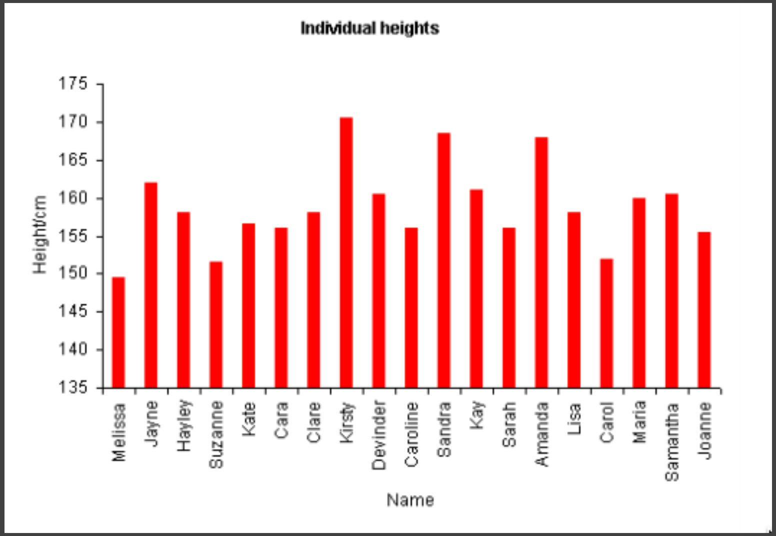


Yearly CO_2 concentrations [Cleveland 85]

Include Zero in Axis Scale?

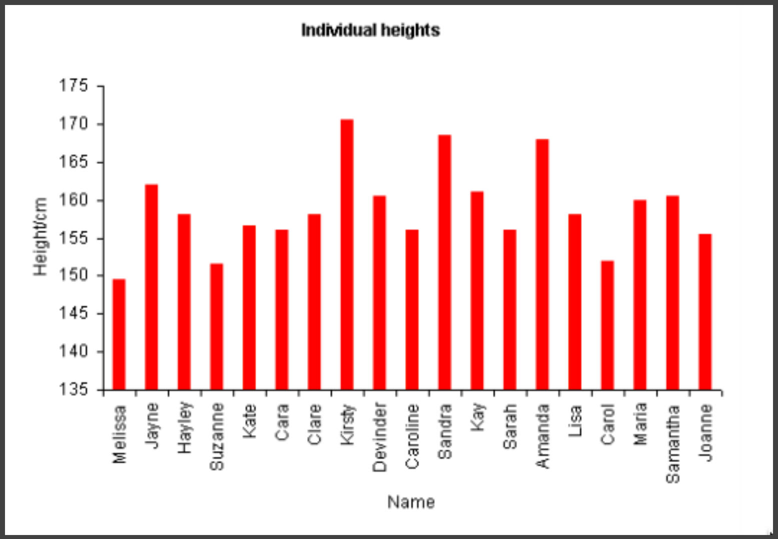


Include Zero in Axis Scale?



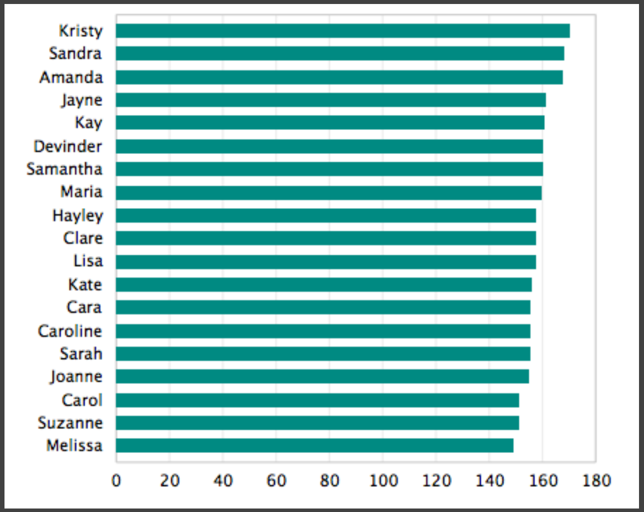
Violates Expressiveness Principle!

Include Zero in Axis Scale?

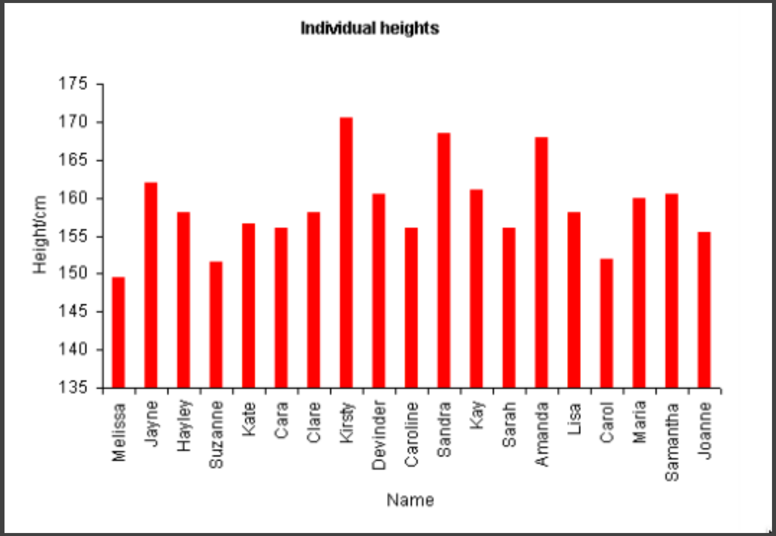


Violates Expressiveness Principle!

Compare Proportions (Q-Ratio)

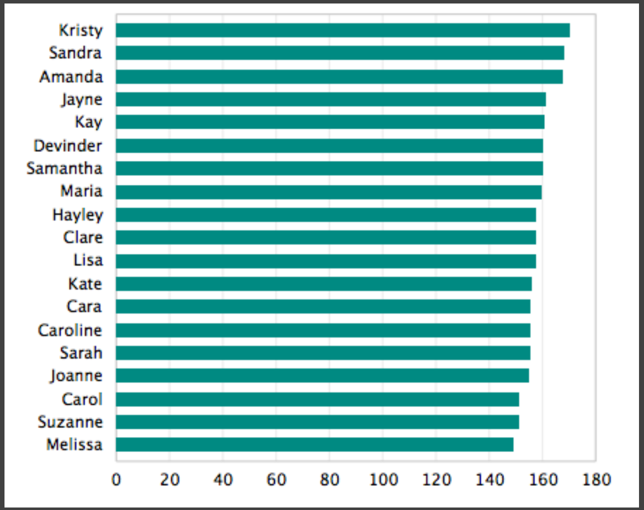


Include Zero in Axis Scale?

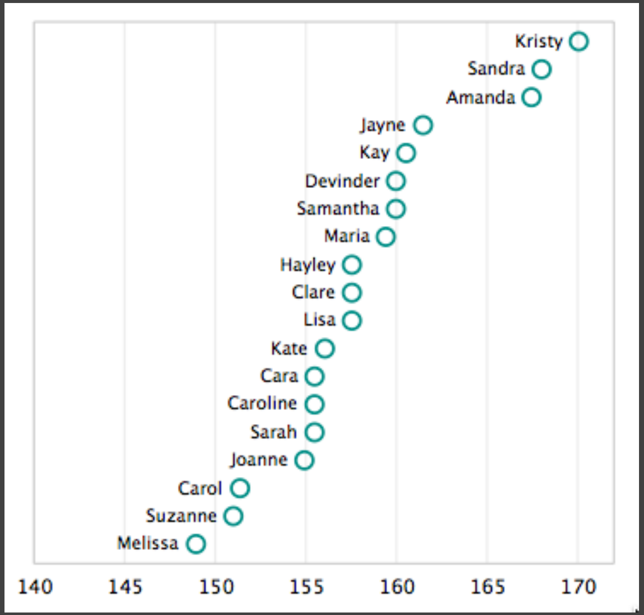


Violates Expressiveness Principle!

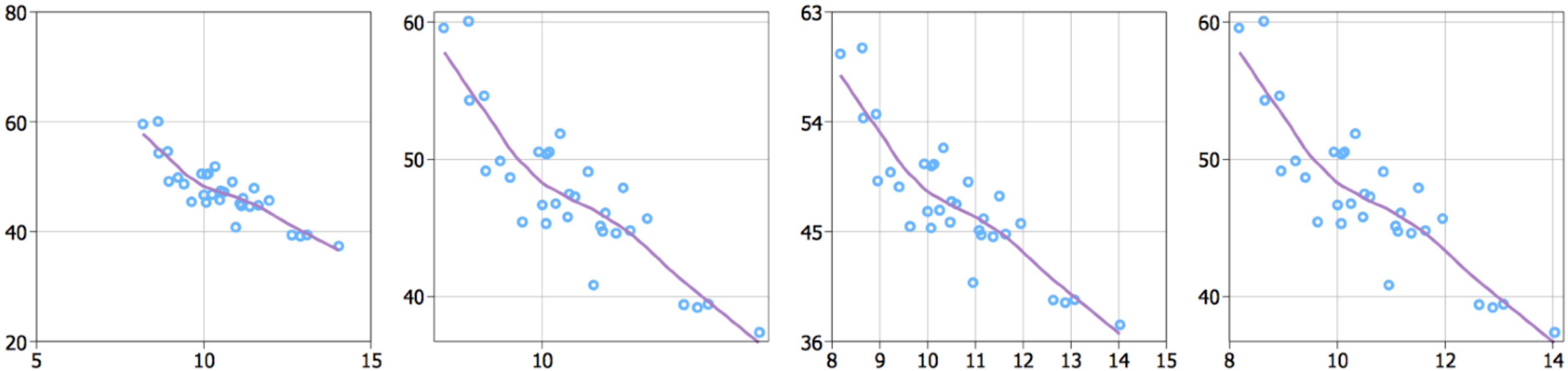
Compare Proportions (Q-Ratio)



Compare Relative Position (Q-Interval)

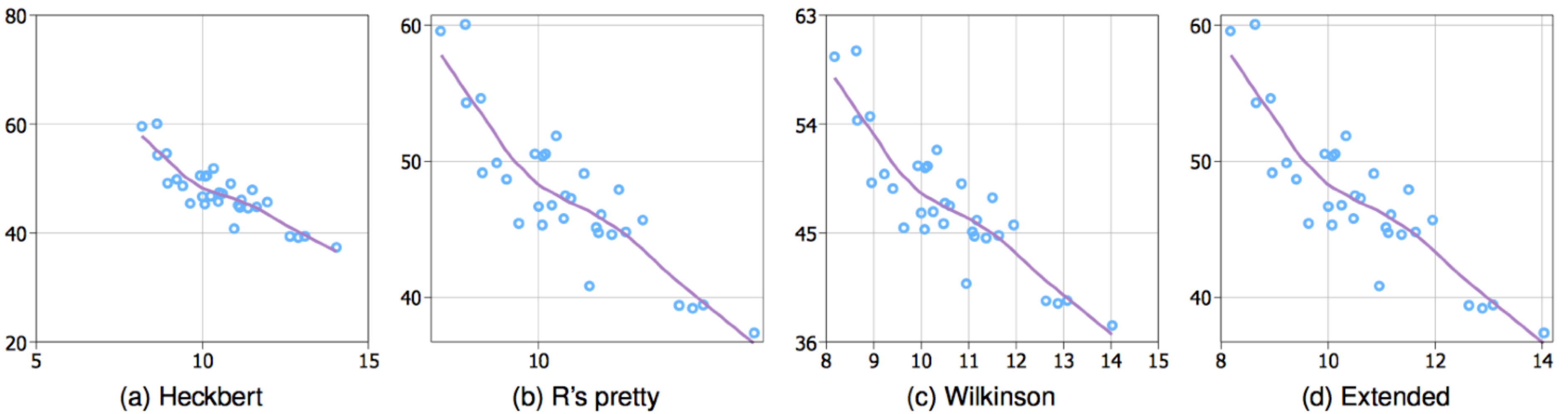


Axis Tick Mark Selection



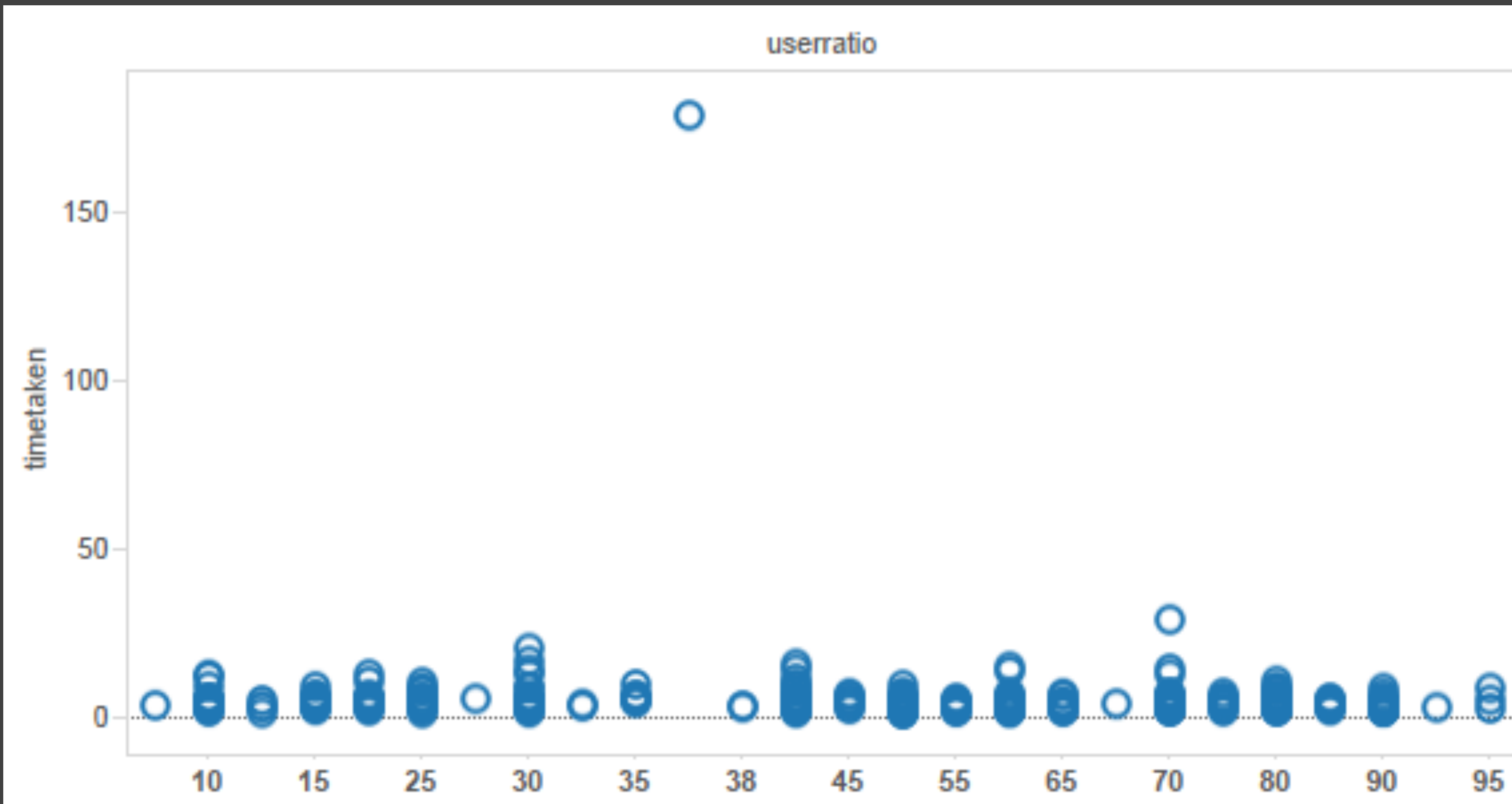
What are some properties of “good” tick marks?

Axis Tick Mark Selection

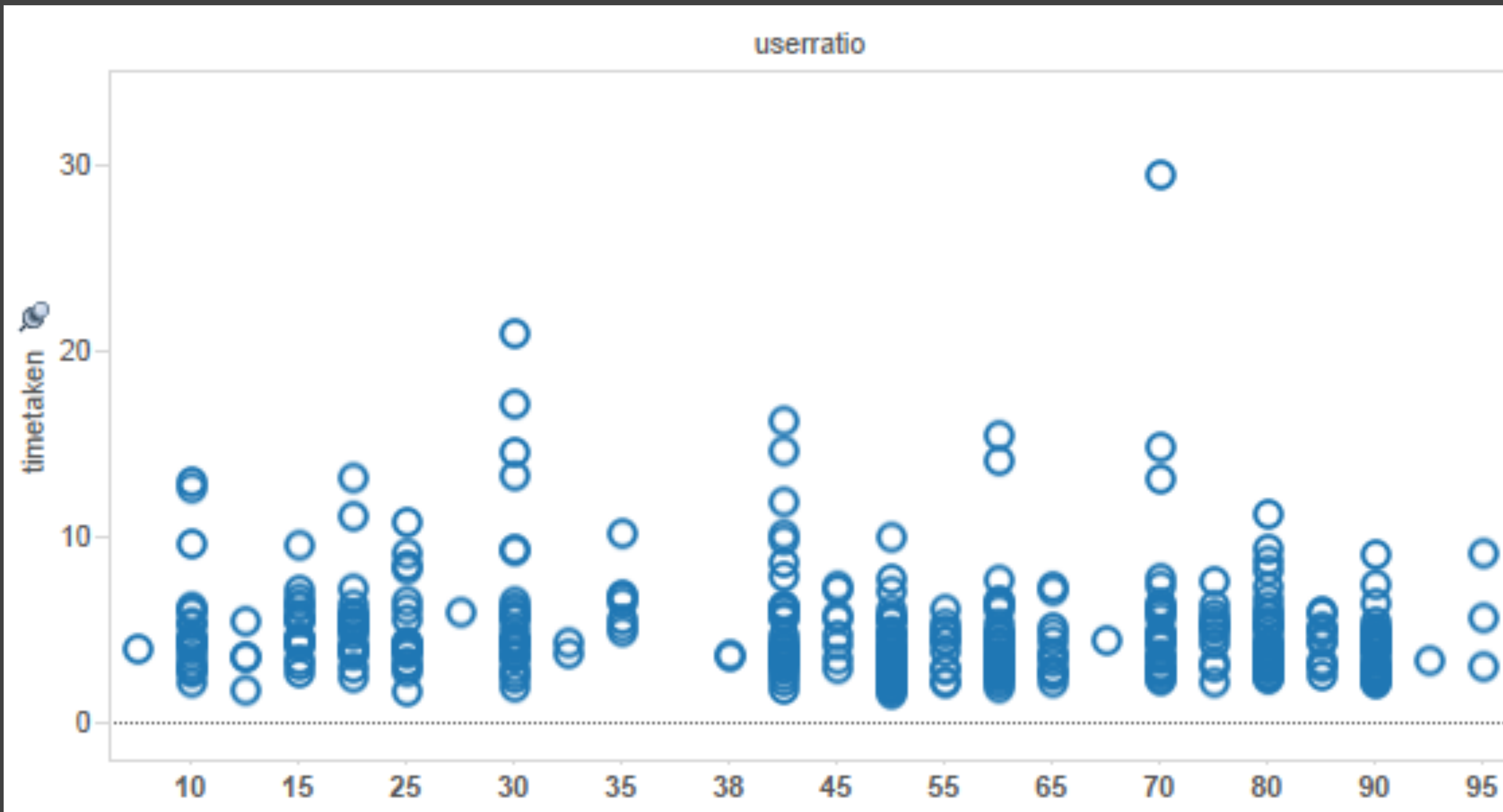


Simplicity - numbers are multiples of 10, 5, 2
Coverage - ticks near the ends of the data
Density - not too many, nor too few
Legibility - whitespace, horizontal text, size

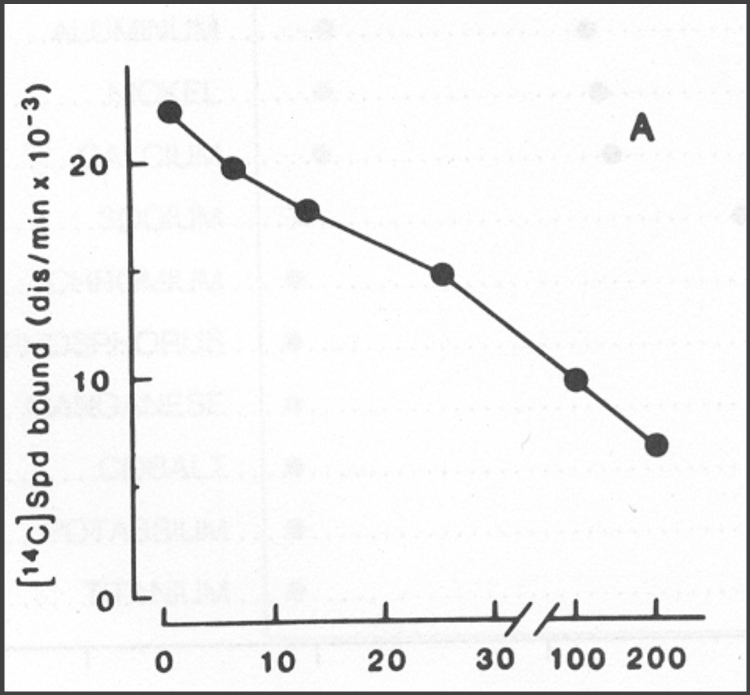
How to Scale the Axis?



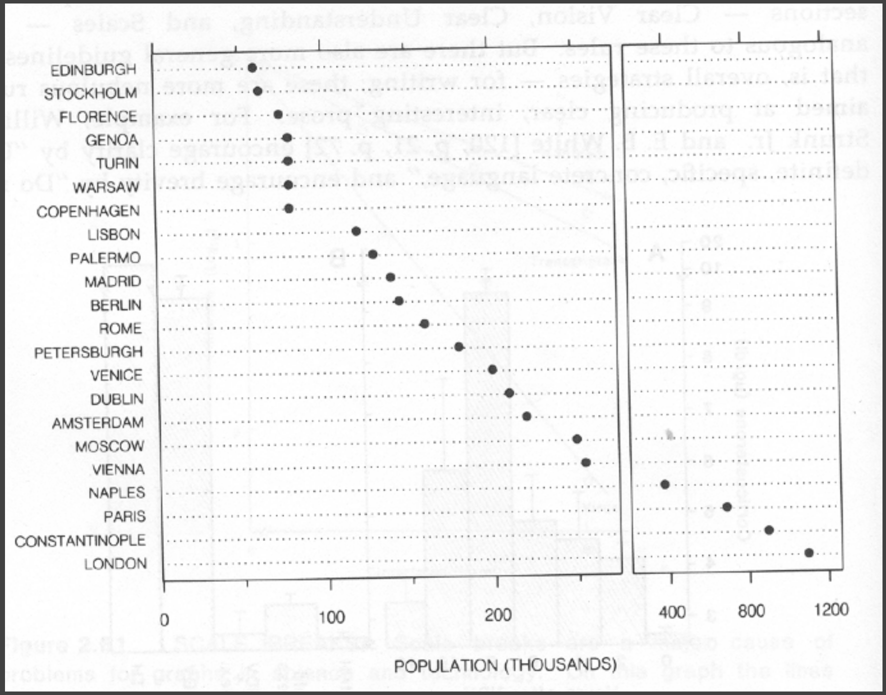
One Option: Clip Outliers



Clearly Mark Scale Breaks



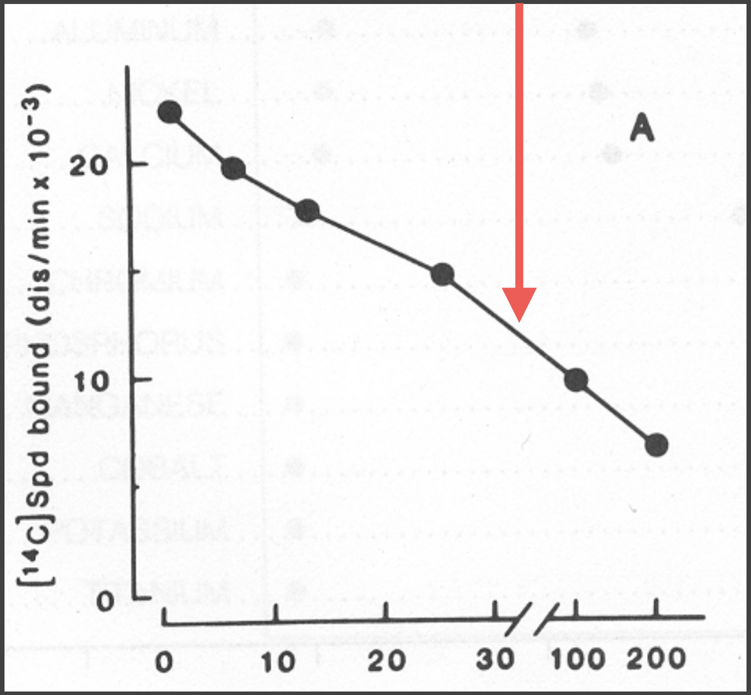
Poor scale break [Cleveland 85]



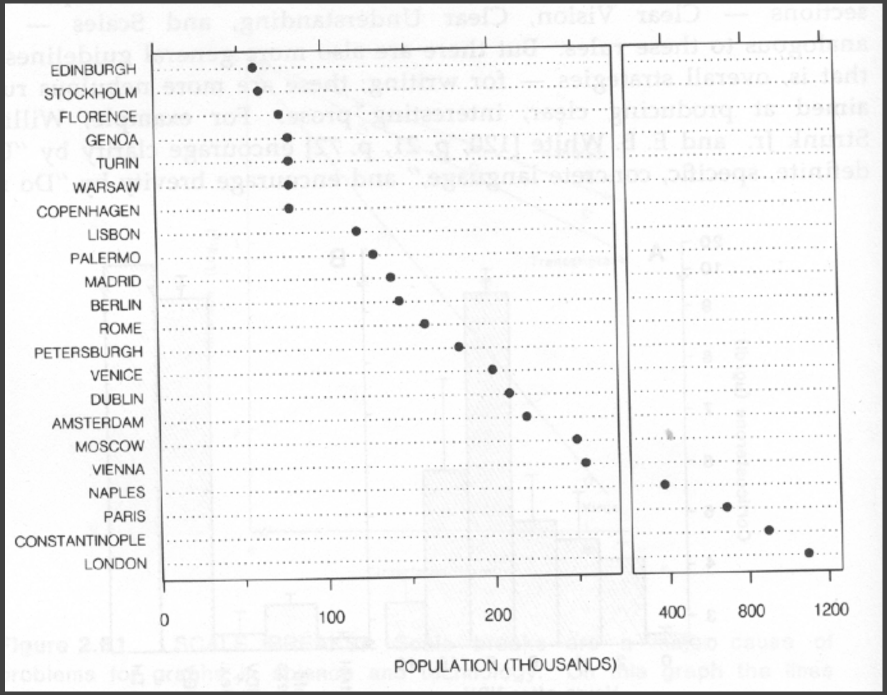
Well-marked scale break [Cleveland 85]

Clearly Mark Scale Breaks

Violates Expressiveness Principle!

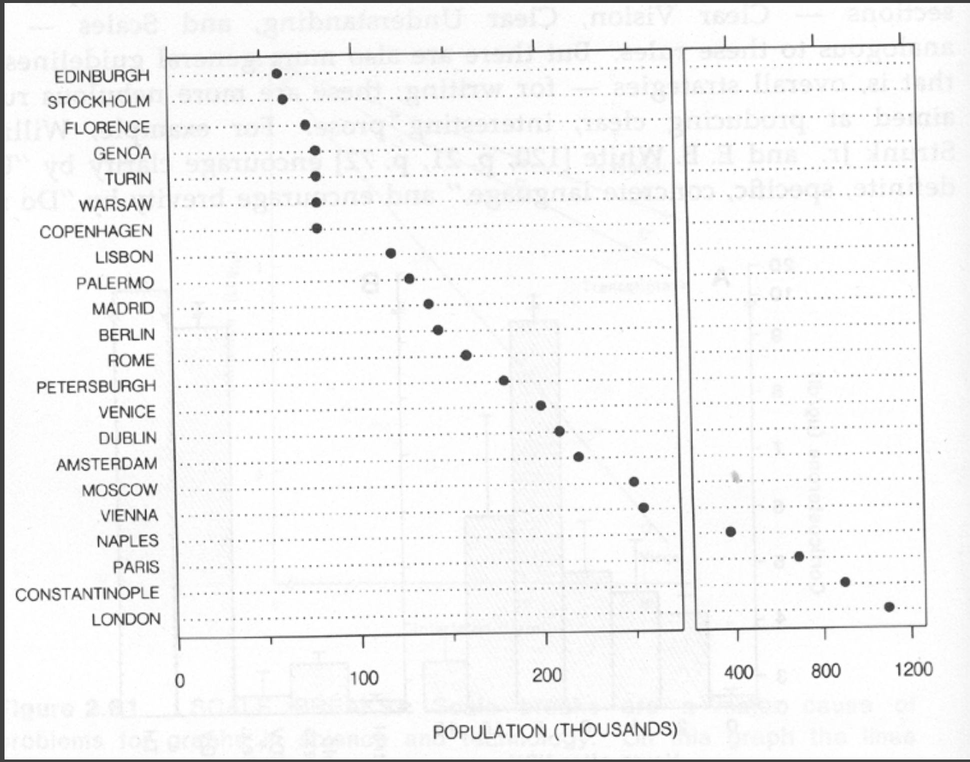


Poor scale break [Cleveland 85]

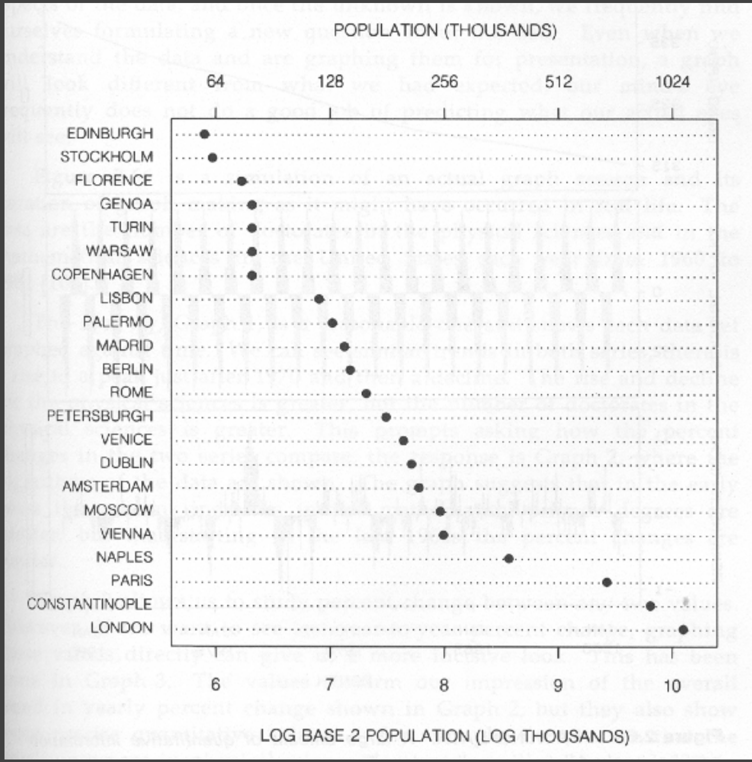


Well-marked scale break [Cleveland 85]

Scale Break vs. Log Scale



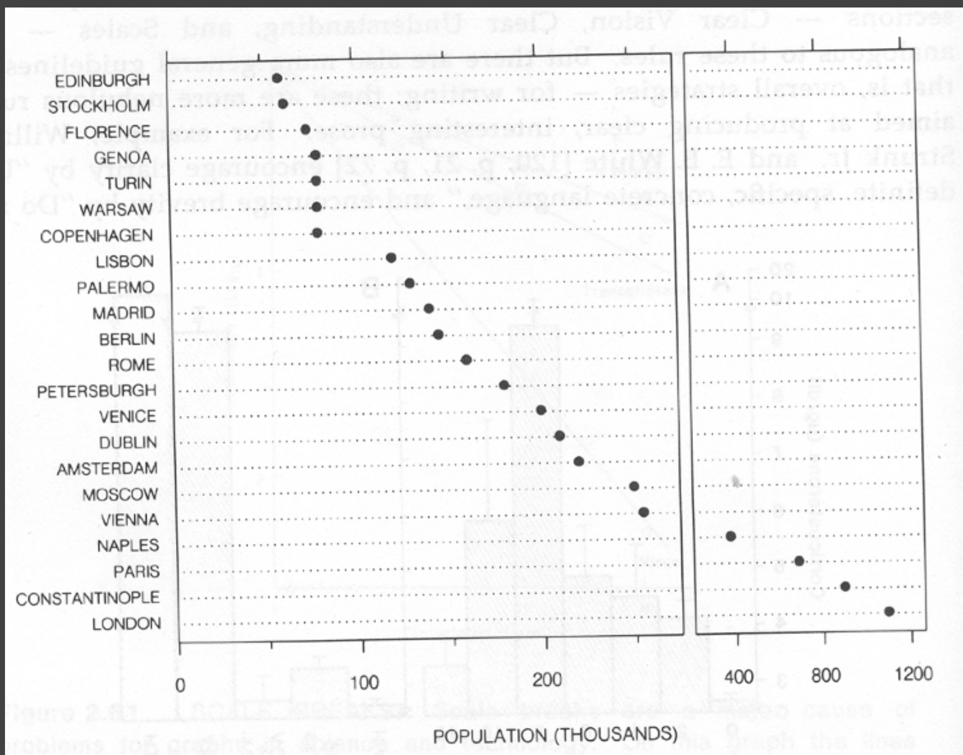
Scale Break

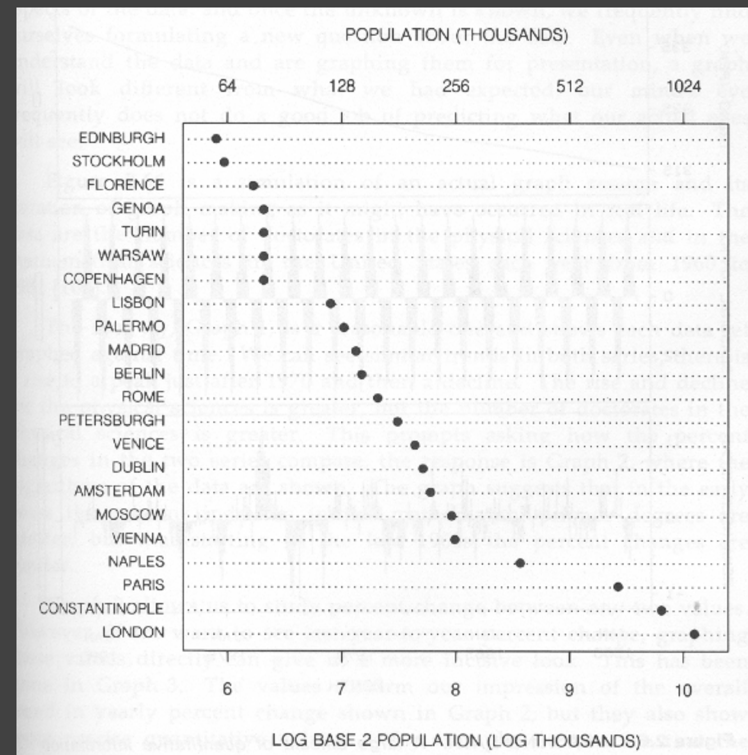


Log Scale

[Cleveland 85]

Scale Break vs. Log Scale





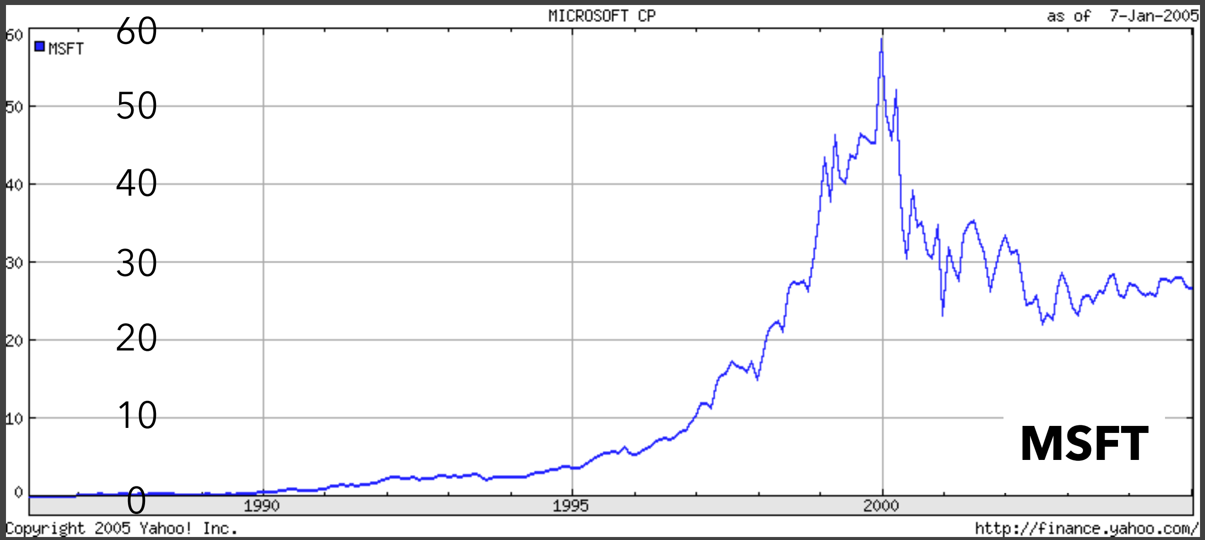
Both increase visual resolution

Scale break: difficult to compare (*cognitive* – not *perceptual* – work)

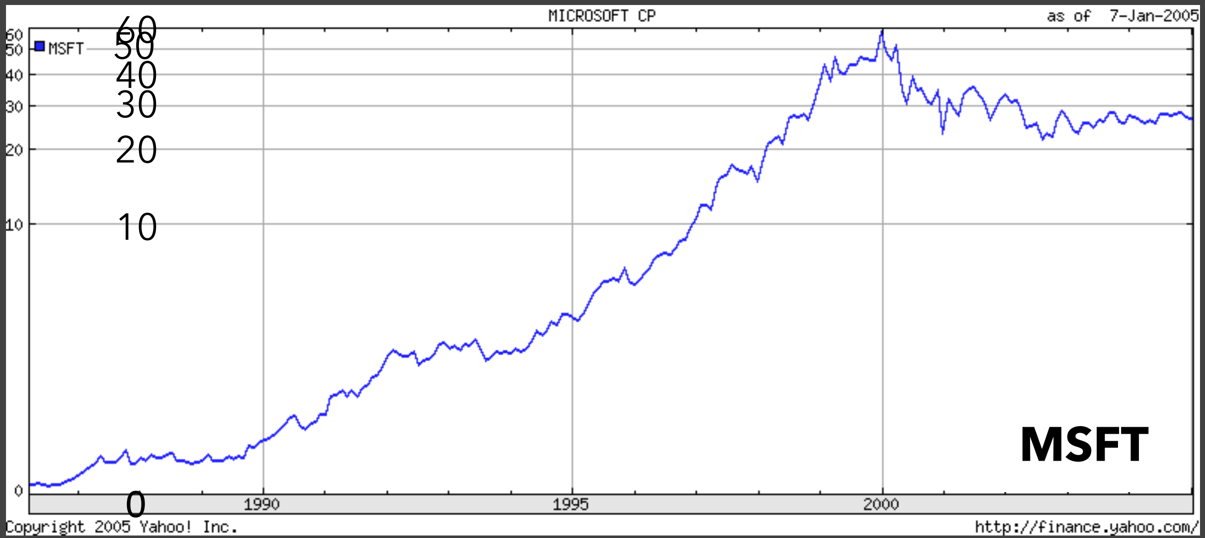
Log scale: direct comparison of all data

Linear Scale vs. Log Scale

Linear Scale



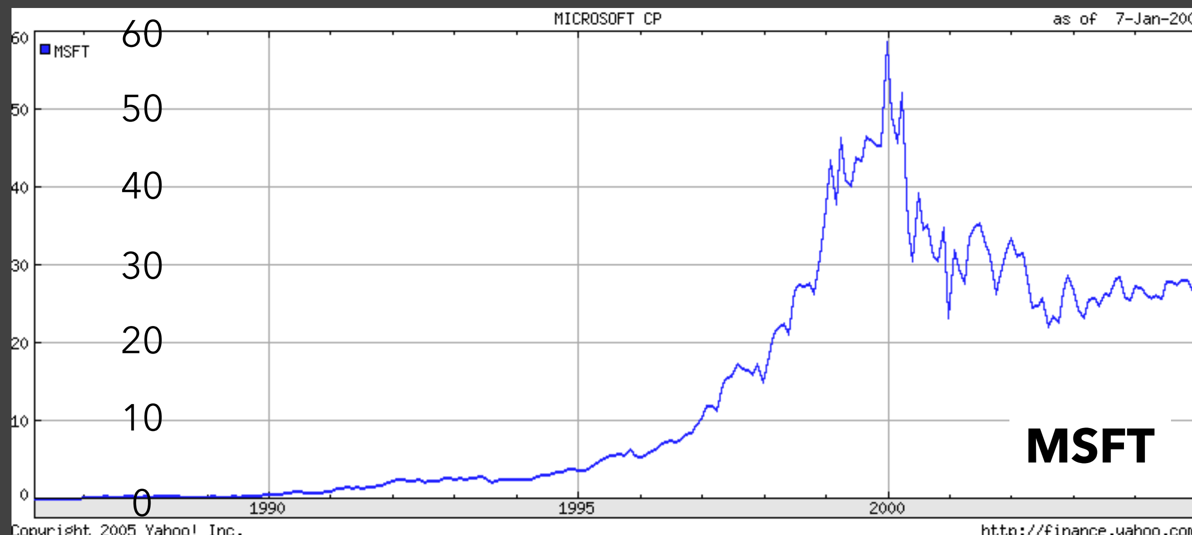
Log Scale



Linear Scale vs. Log Scale

Linear Scale

Absolute change

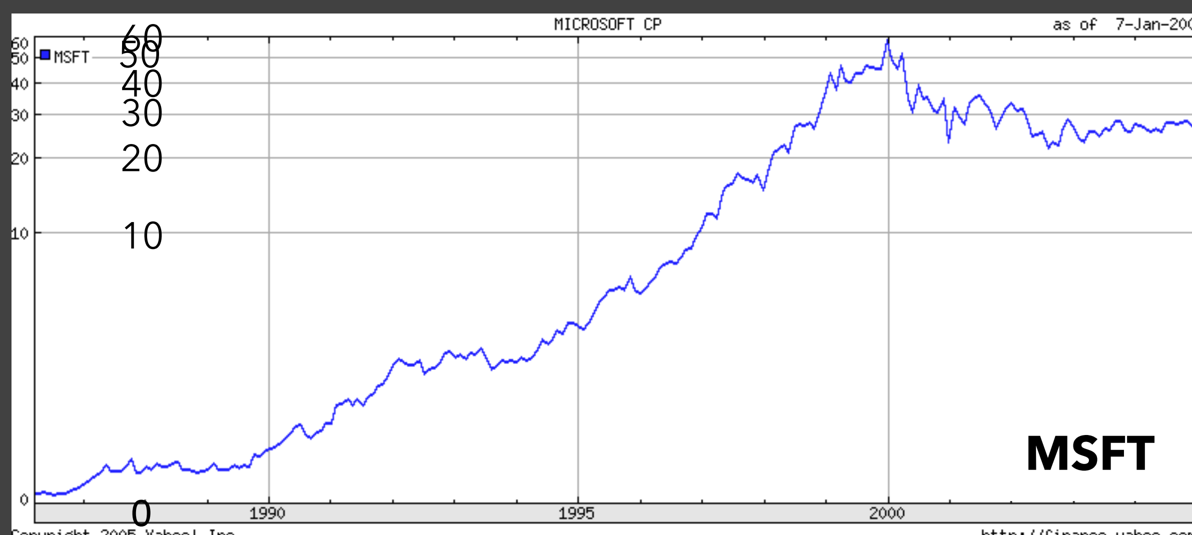


Log Scale

Small fluctuations

Percent change

d(10,20) = d(30,60)



When To Apply a Log Scale?

Address data skew (e.g., long tails, outliers)

Enables comparison within and across multiple orders of magnitude.

Focus on multiplicative factors (not additive)

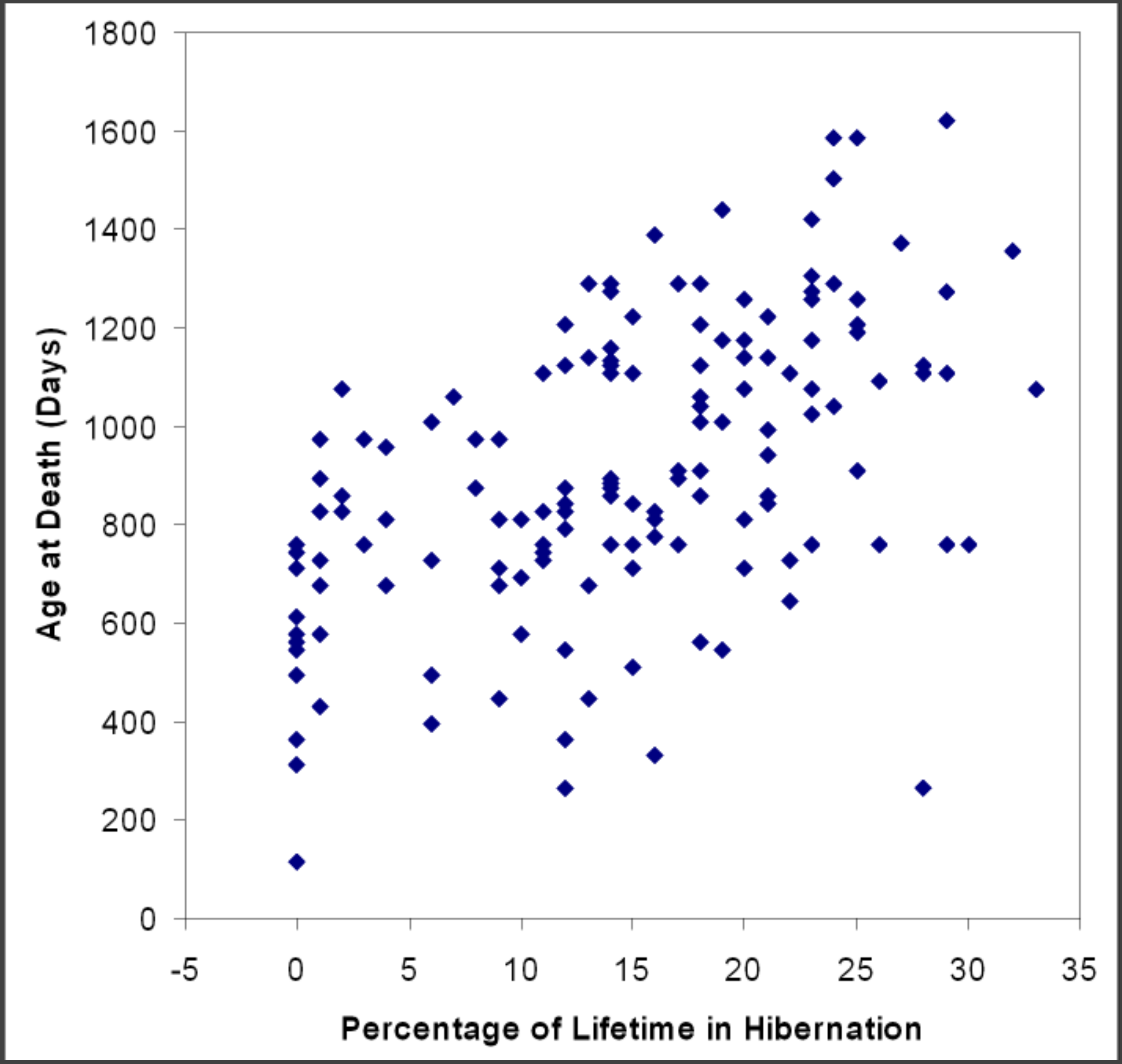
Recall that the logarithm transforms **×** to **+**!

Percentage change, not absolute value.

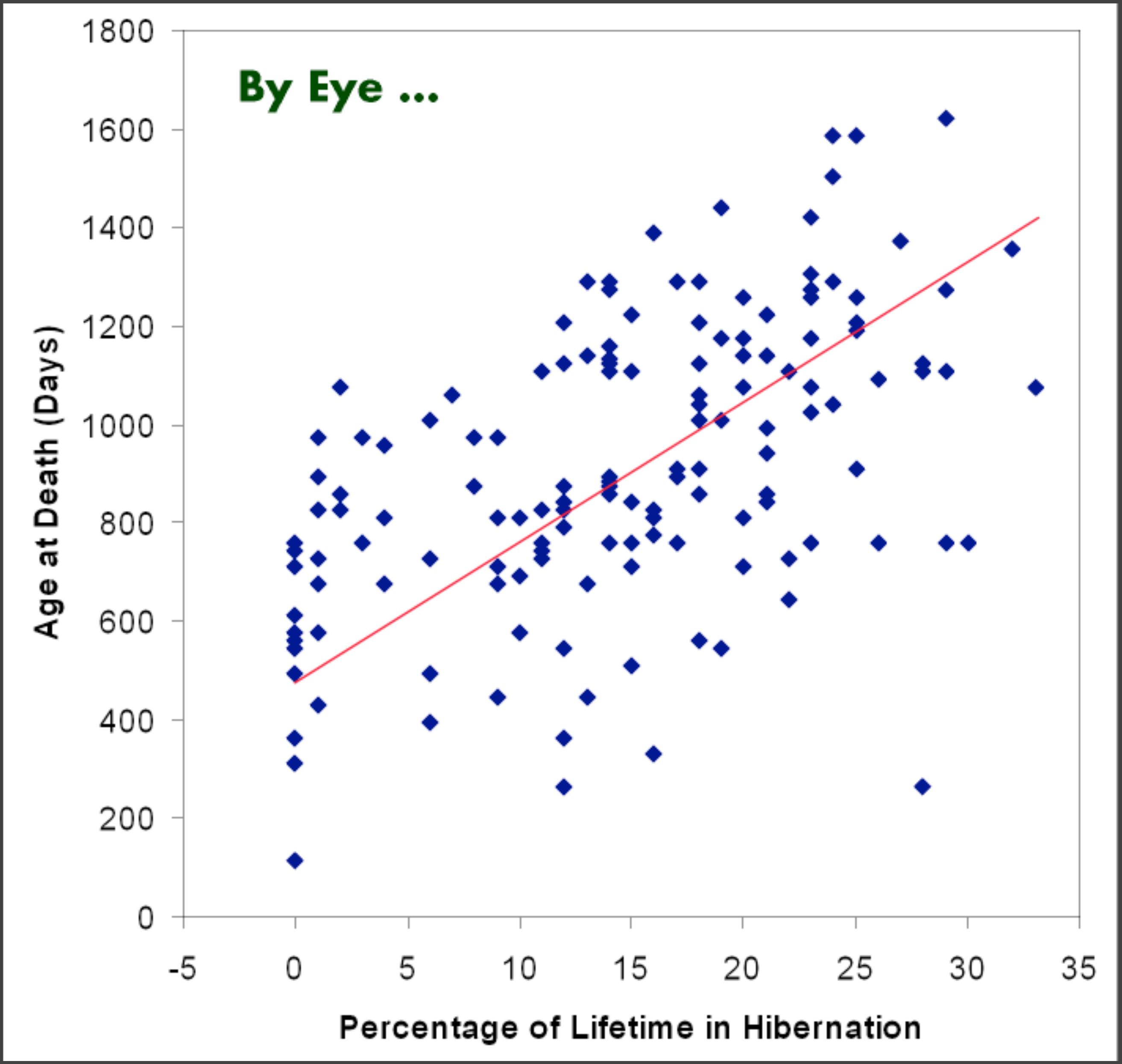
Constraint: **positive, non-zero values**

Constraint: **audience familiarity?**

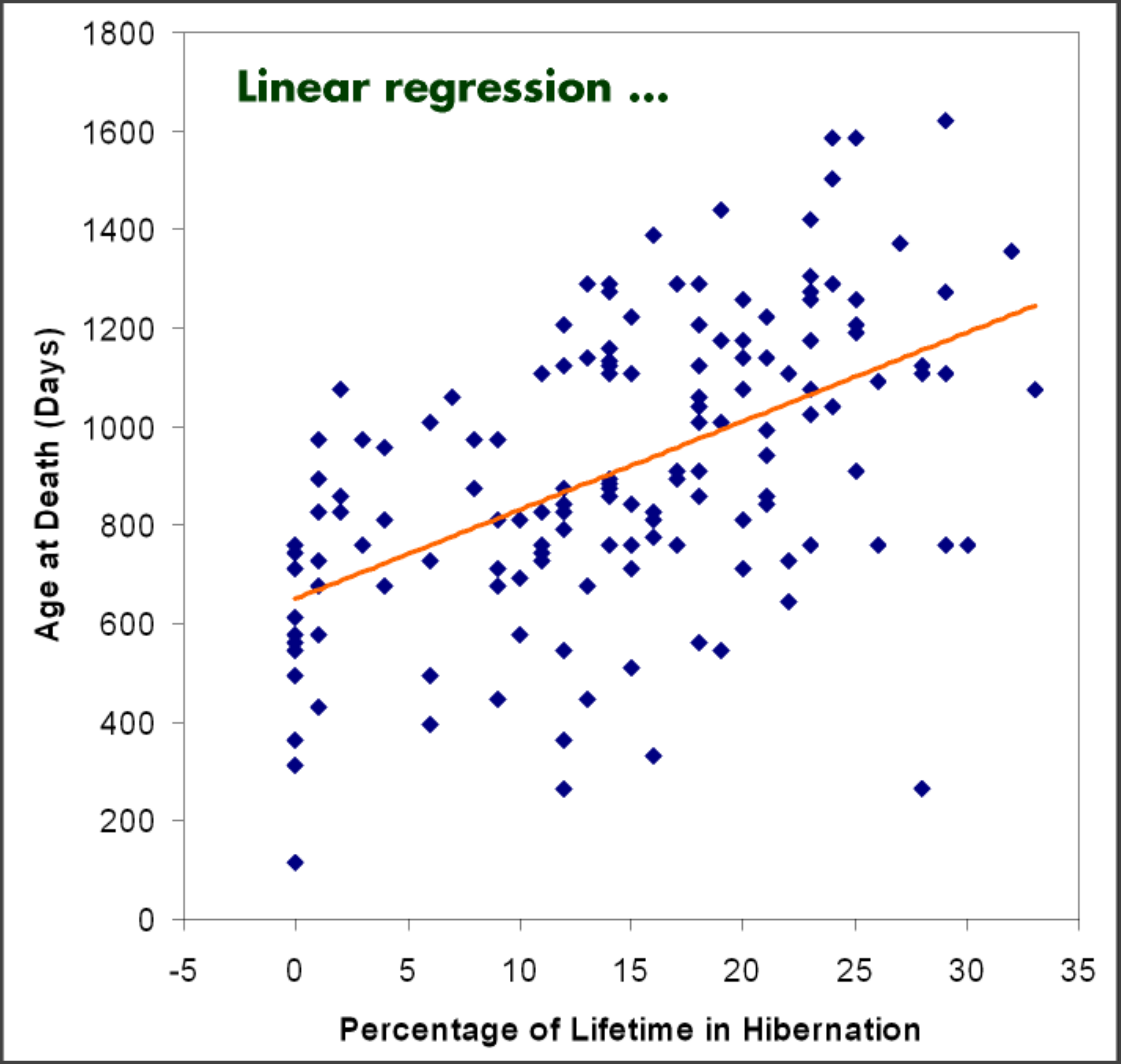
Regression Lines



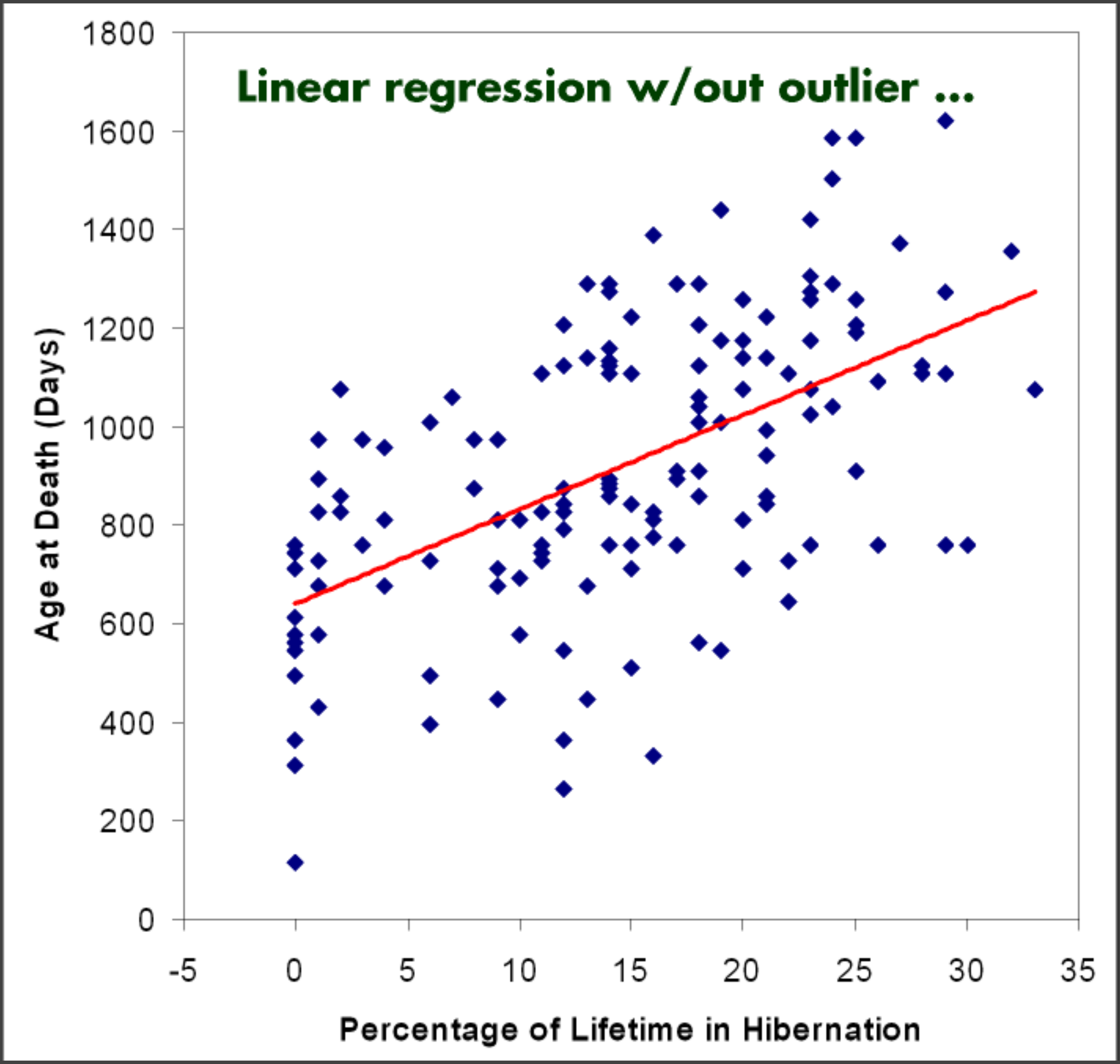
[The Elements of Graphing Data. Cleveland 94]



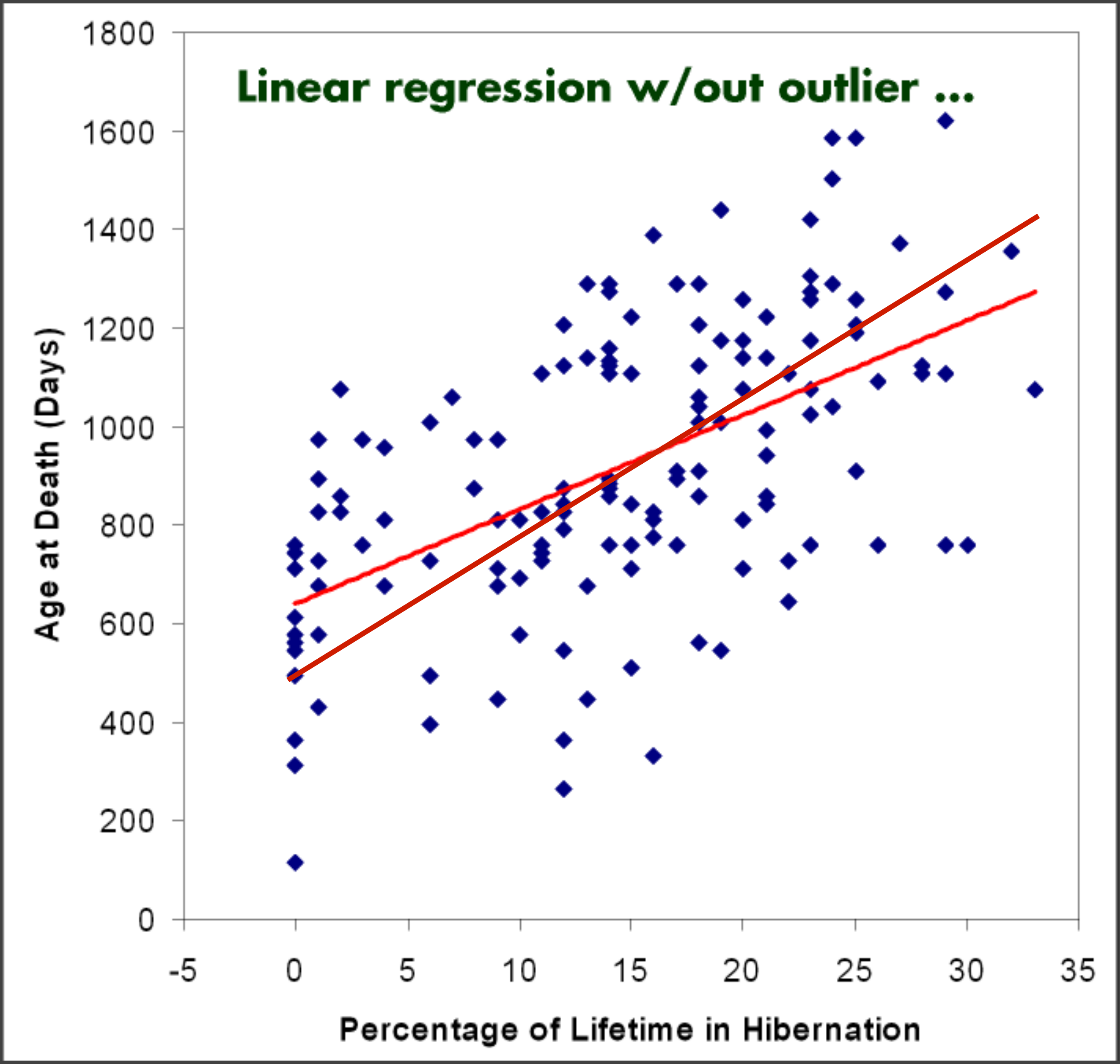
[The Elements of Graphing Data. Cleveland 94]



[The Elements of Graphing Data. Cleveland 94]



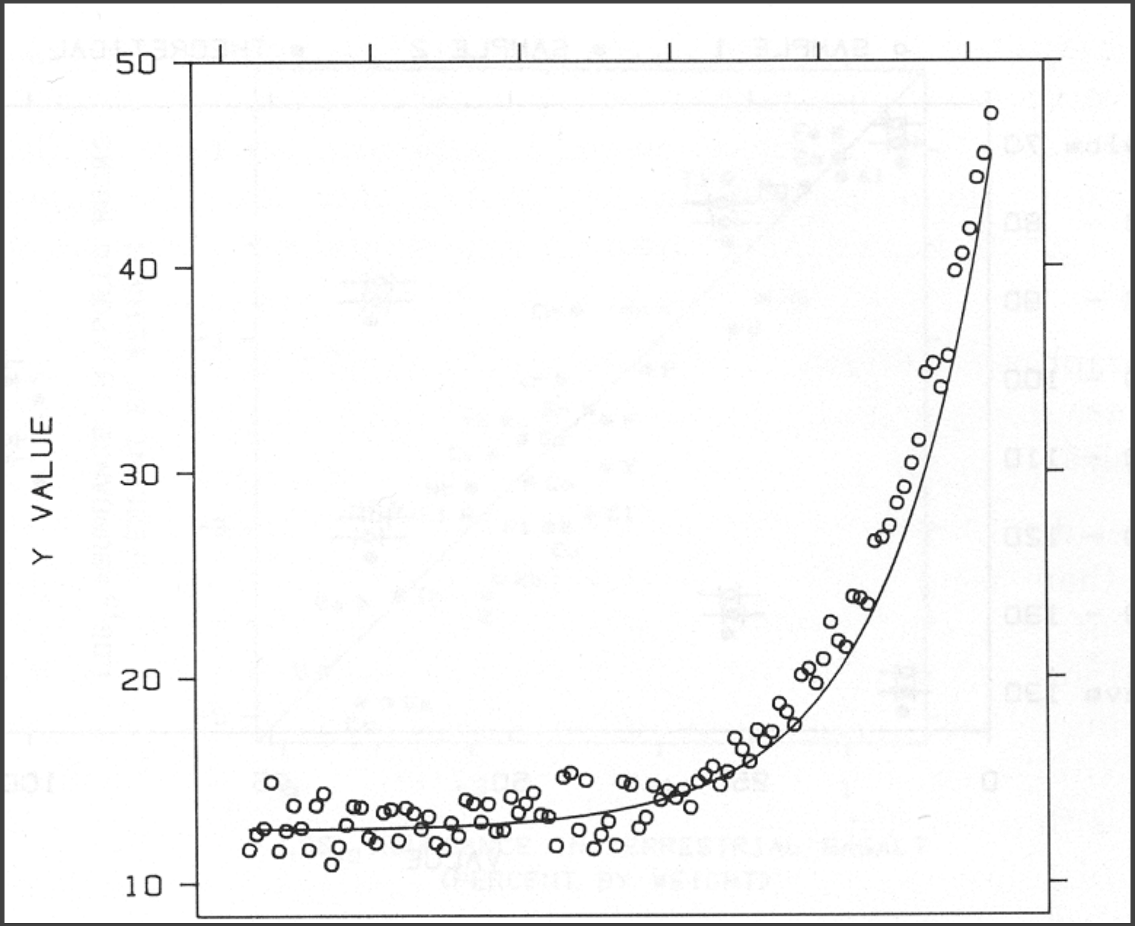
[The Elements of Graphing Data. Cleveland 94]



[The Elements of Graphing Data. Cleveland 94]

Transforming Data

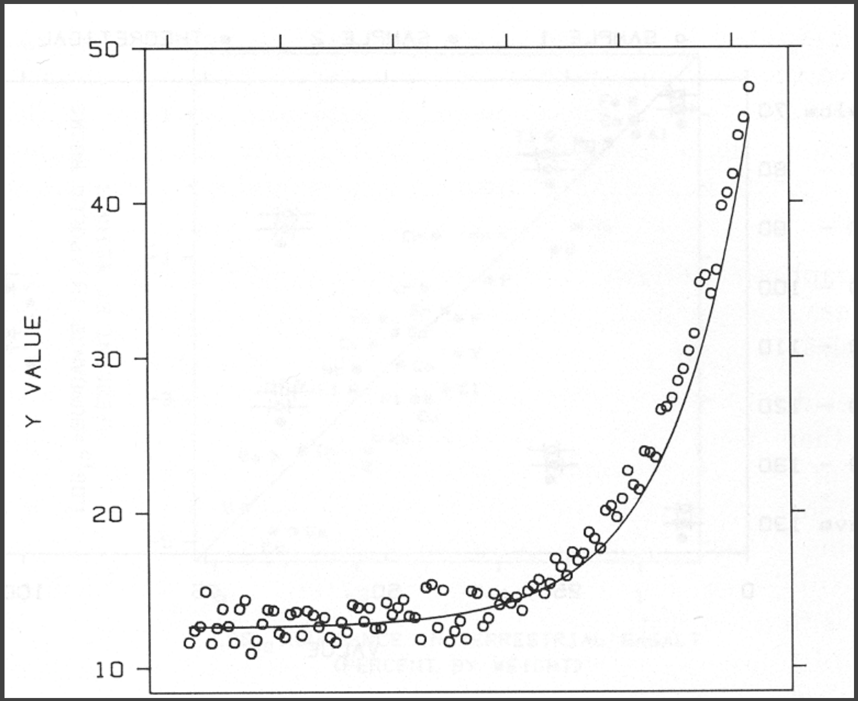
How well does the curve fit the data?

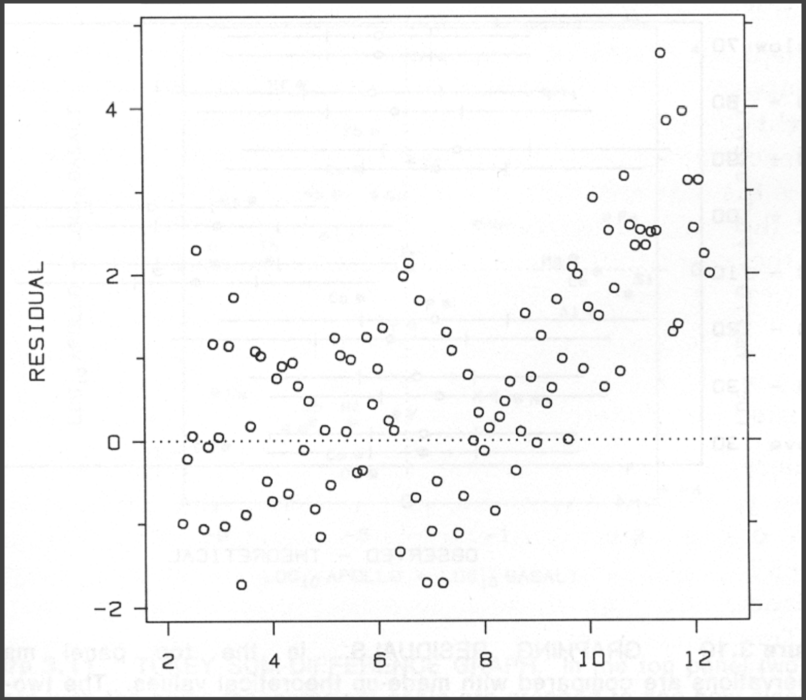


[Cleveland 85]

Plot the Residuals

Plot vertical distance from best fit curve
Residual graph shows accuracy of fit

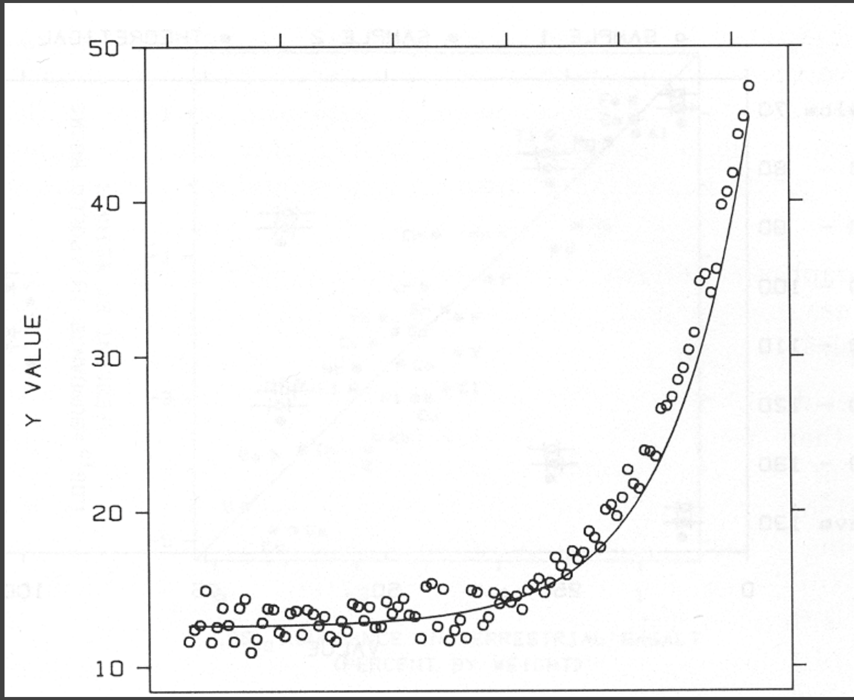




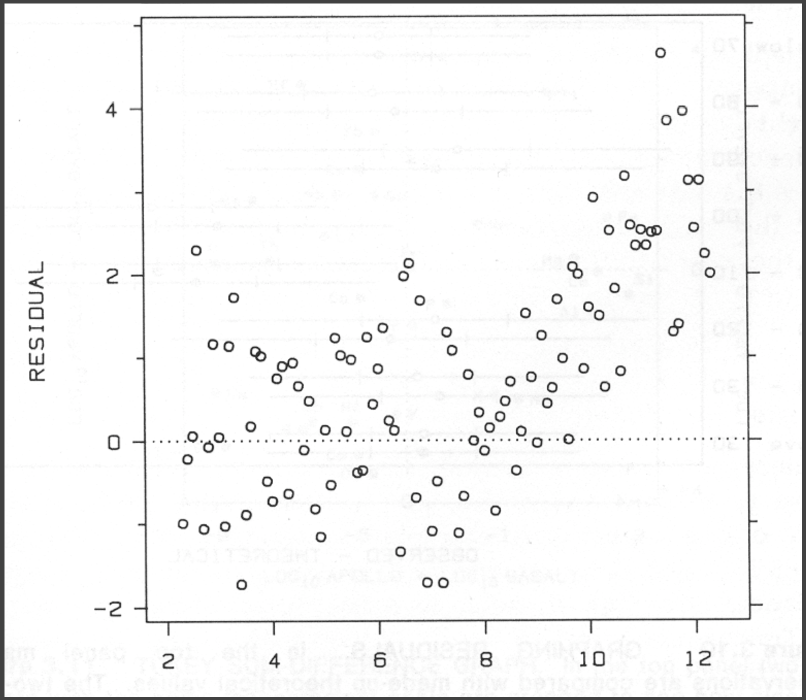
[Cleveland 85]

Multiple Plotting Options

Plot model in data space



Plot data in model space



[Cleveland 85]

Administrivia

A2: Exploratory Data Analysis

Use visualization software to form & answer questions

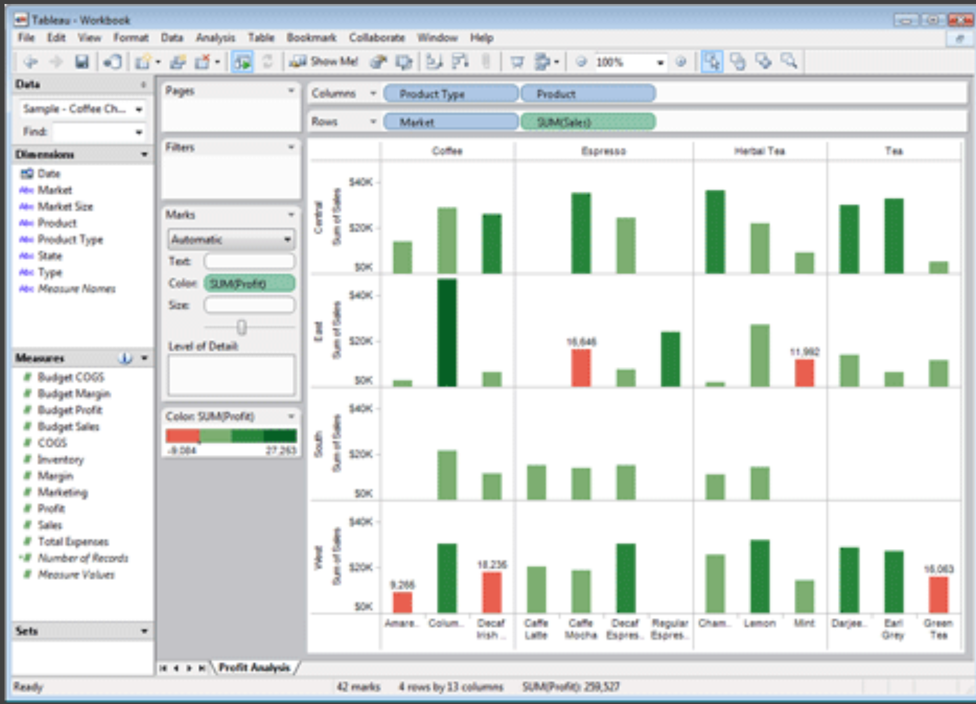
First steps:

Step 1: Pick domain & data

Step 2: Pose questions

Step 3: Profile the data

Iterate as needed



Create visualizations

Interact with data

Refine your questions

Author a report

Screenshots of most insightful views *(10+)*

Include titles and captions for each view

Due by 11:59pm

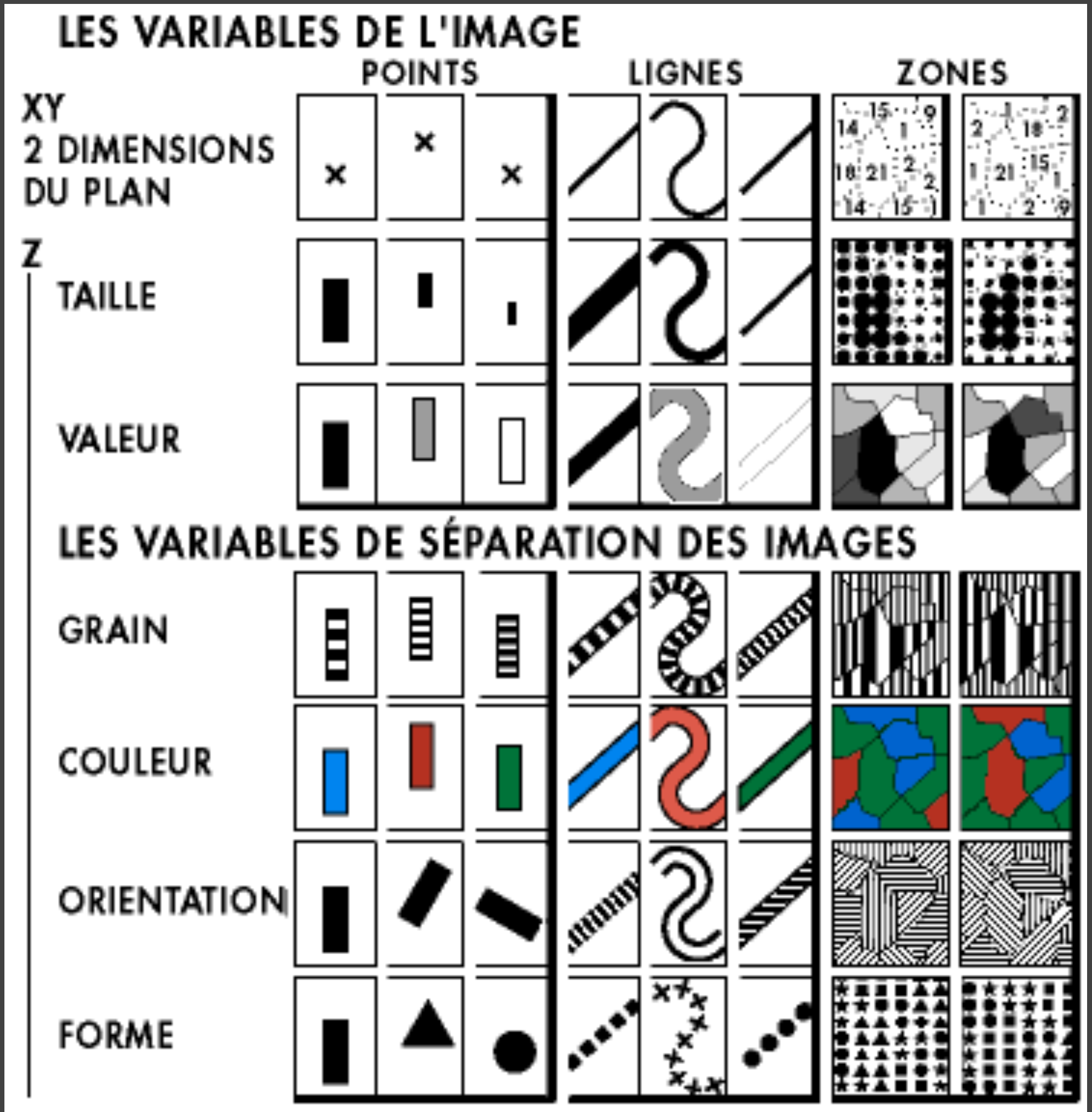
Tuesday, Oct 16

Multidimensional Data

Visual Encoding Variables

Position (X)
Position (Y)
Size
Value
Texture
Color
Orientation
Shape

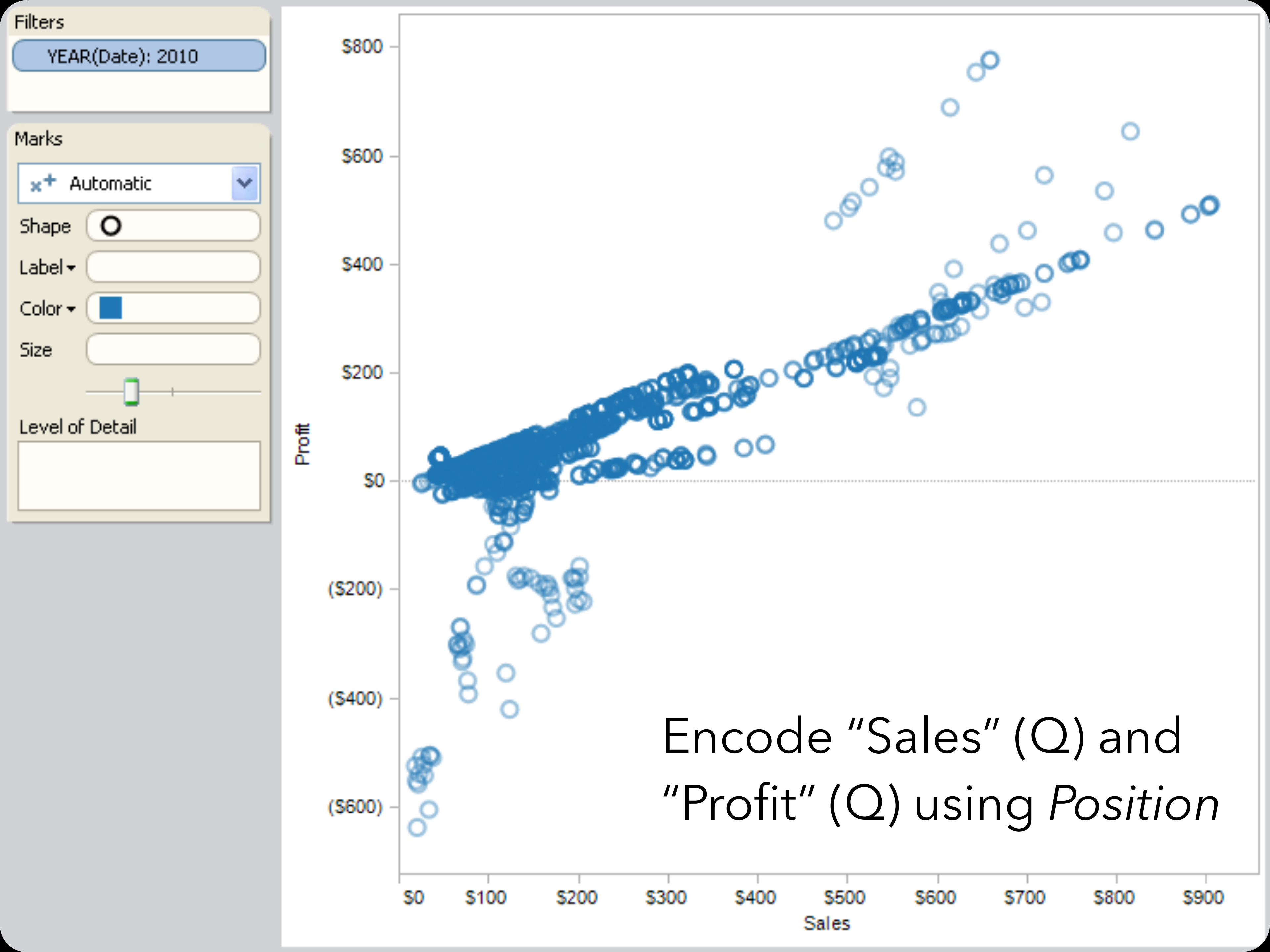
~8 dimensions?



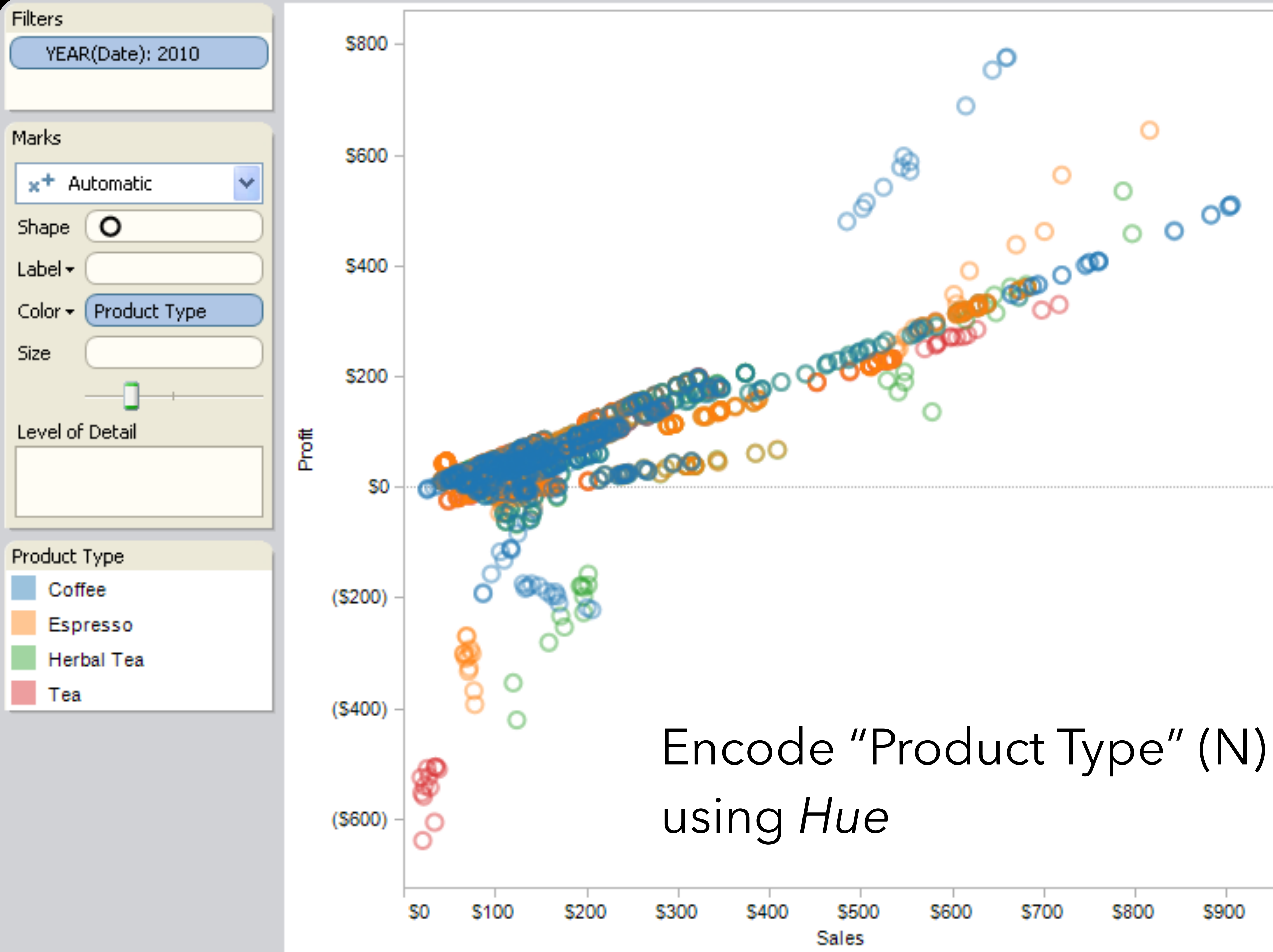
Example: Coffee Sales

Sales figures for a fictional coffee chain

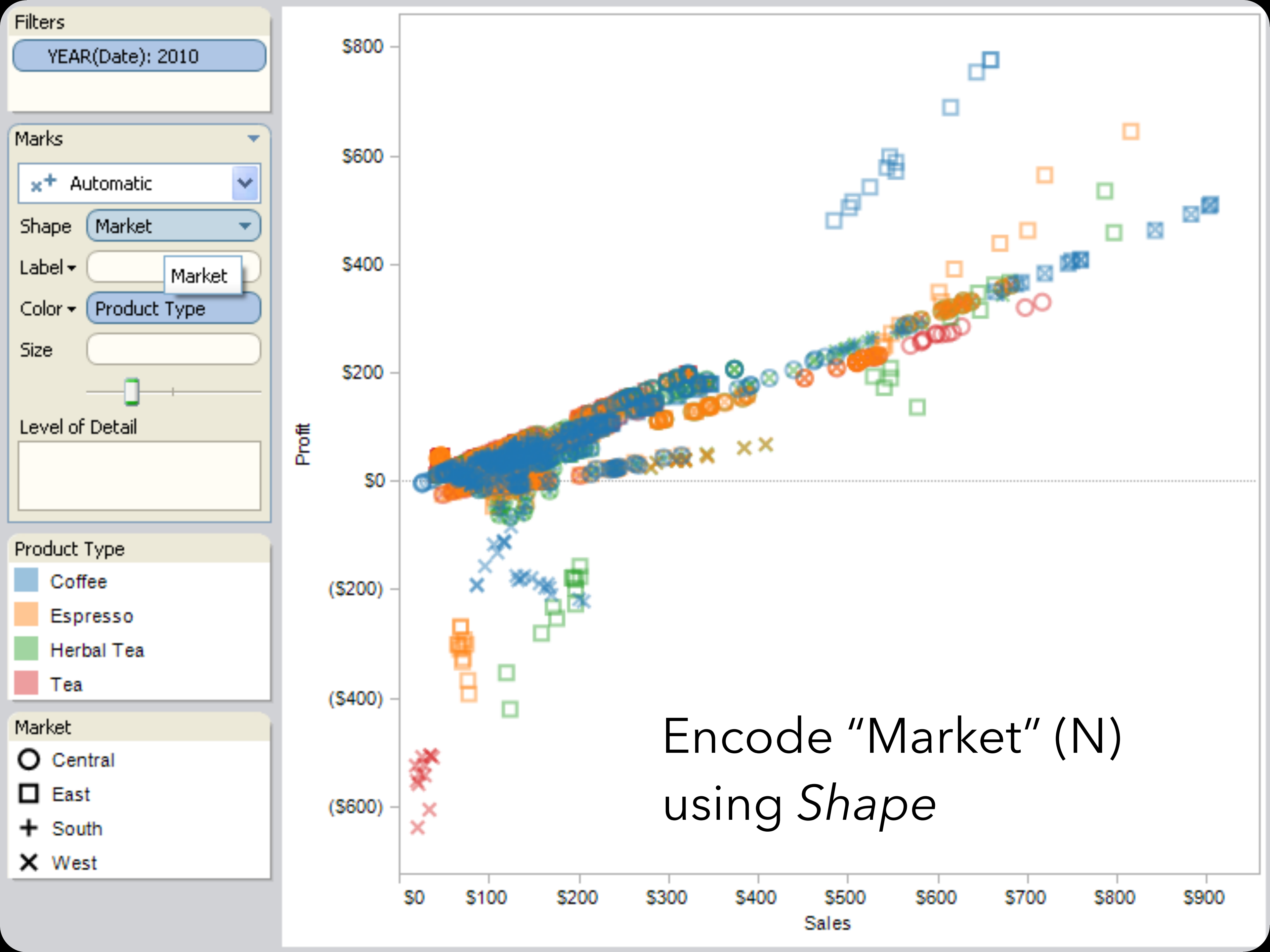
Sales	Q-Ratio
Profit	Q-Ratio
Marketing	Q-Ratio
Product Type	N {Coffee, Espresso, Herbal Tea, Tea}
Market	N {Central, East, South, West}



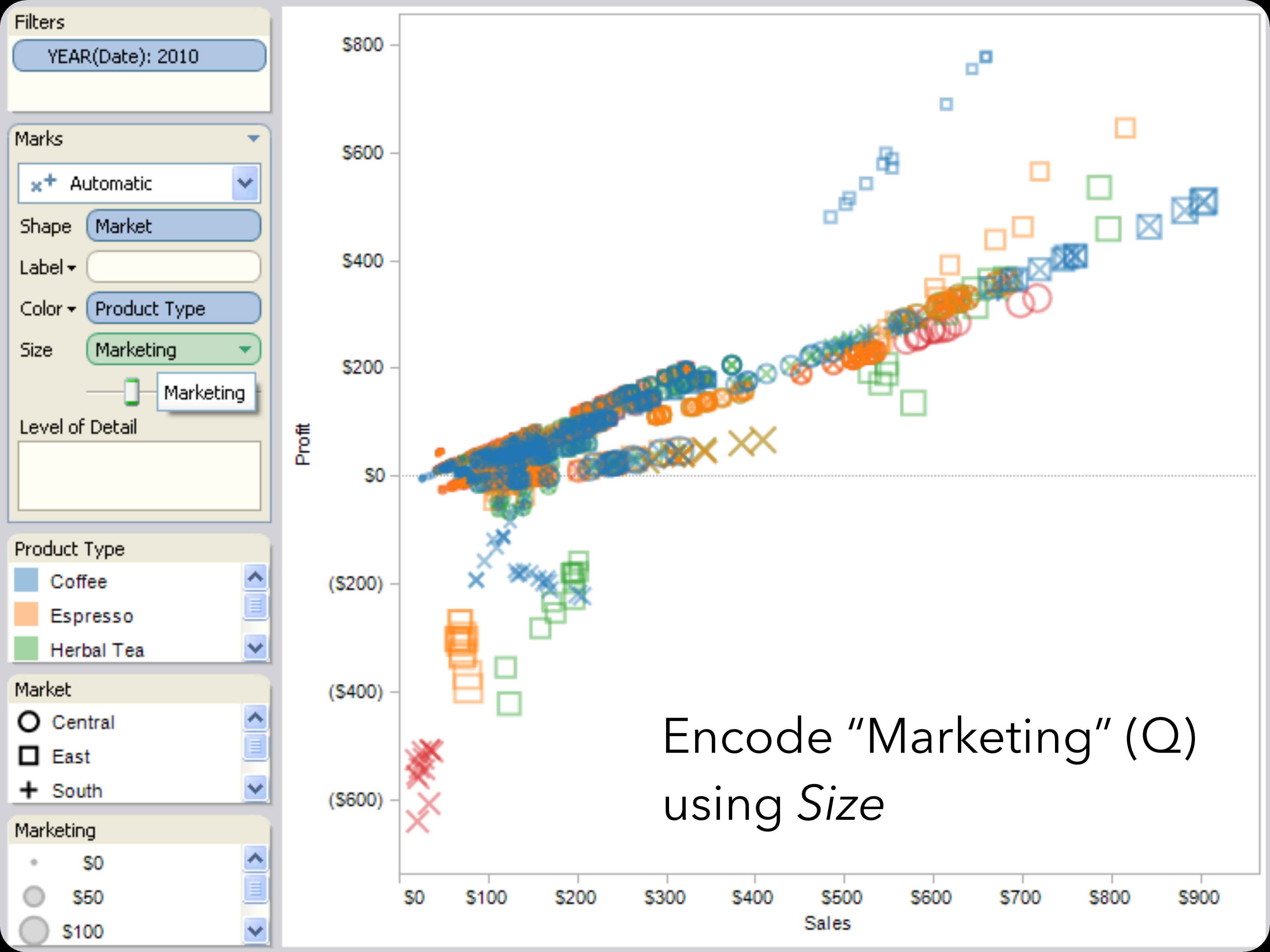
Encode "Sales" (Q) and "Profit" (Q) using *Position*



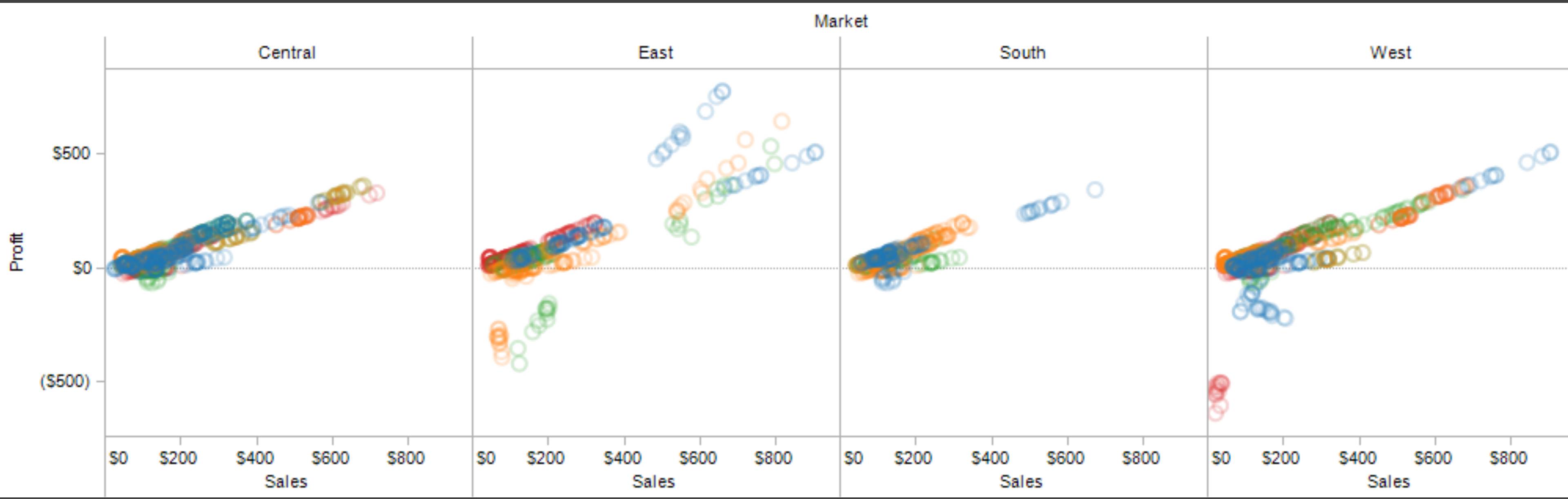
Encode "Product Type" (N) using *Hue*



Encode "Market" (N) using *Shape*


Filters
YEAR(Date): 2010
Marks
Automatic
Shape
Market
Label
Color
Product Type
Size
Marketing
Marketing
Level of Detail
Product Type
Coffee
Espresso
Herbal Tea
Market
Central
East
South
Marketing
$0
$50
$100
Profit
$800
$600
$400
$200
$0
($200)
($400)
($600)
Sales
$0
$100
$200
$300
$400
$500
$600
$700
$800
$900
Encode "Marketing" (Q)
using *Size*

Trellis Plots



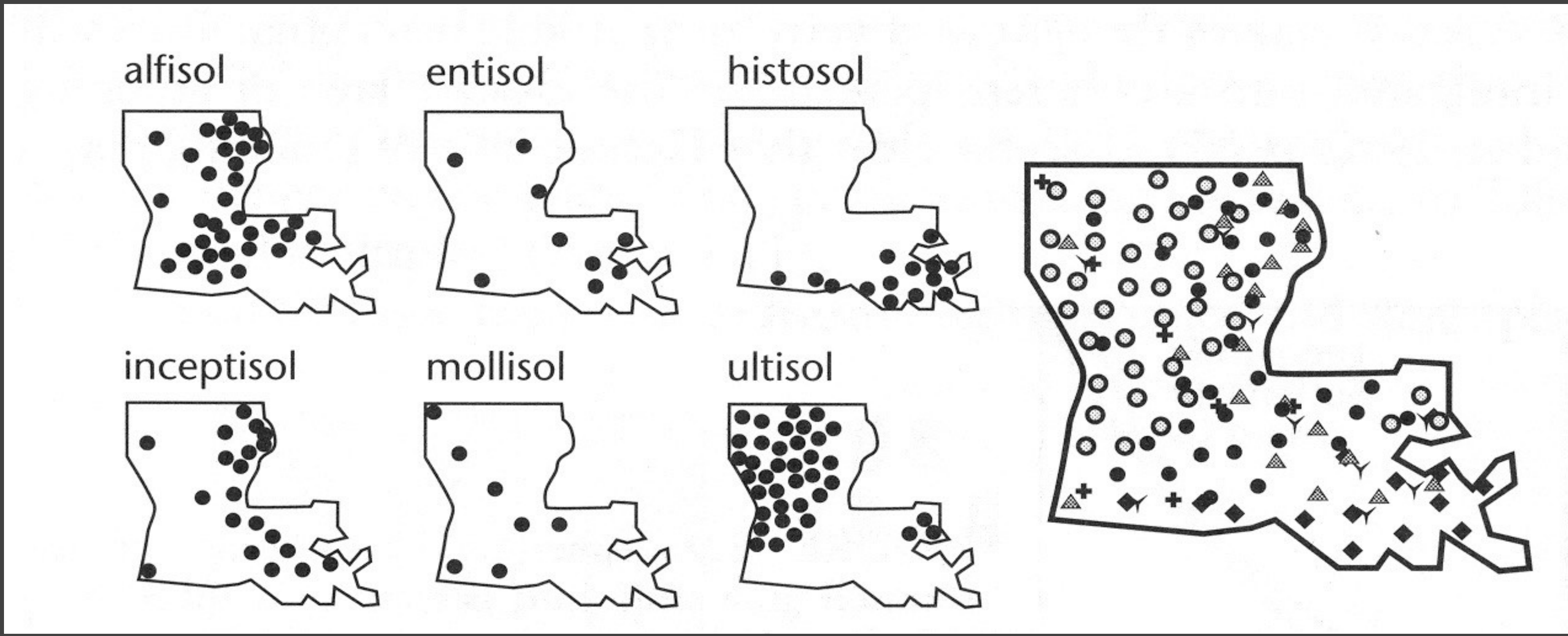
A *trellis plot* subdivides space to enable comparison across multiple plots.

Typically nominal or ordinal variables are used as dimensions for subdivision.

Small Multiples

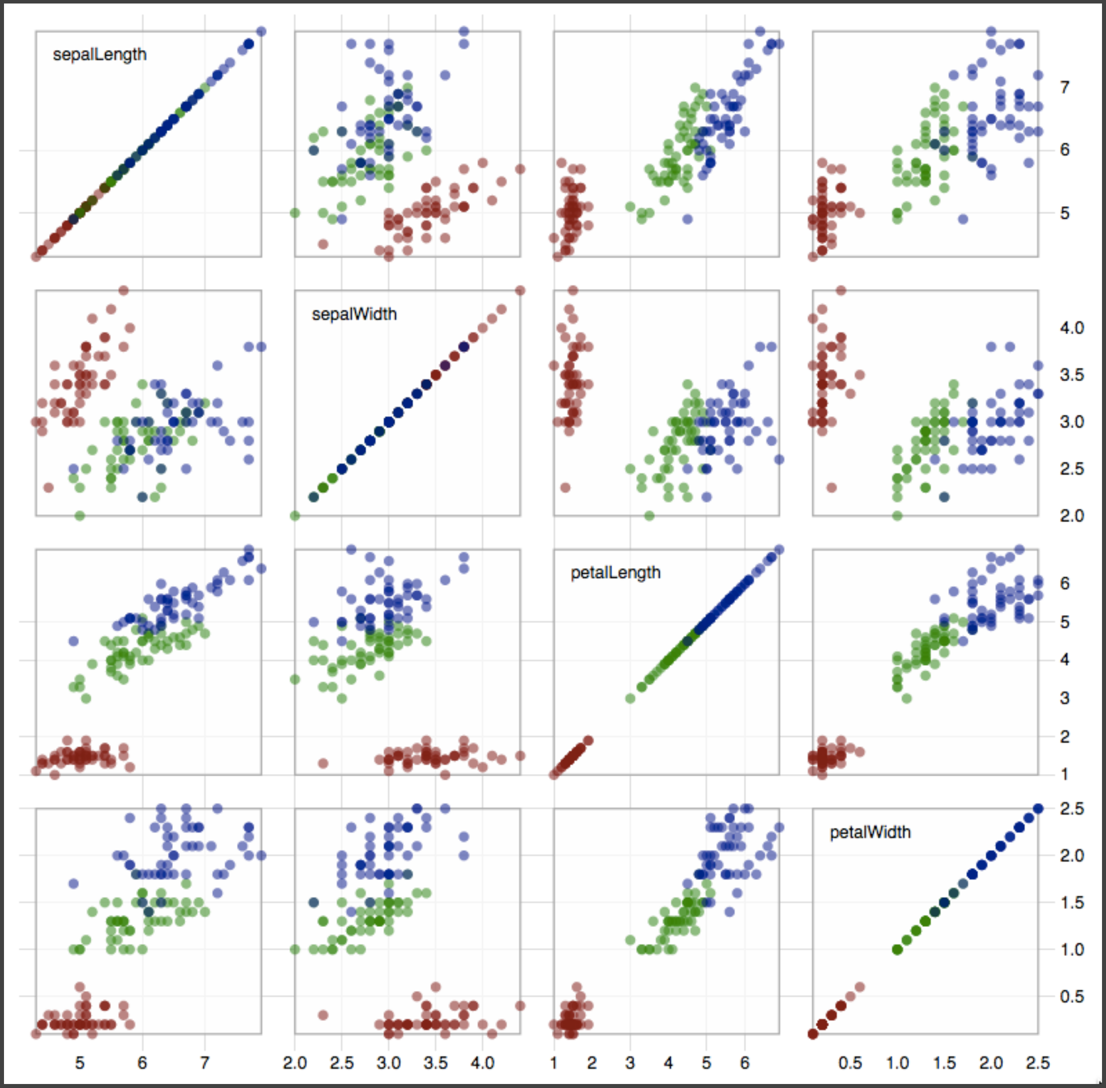
[MacEachren '95, Figure 2.11, p. 38]

Small Multiples



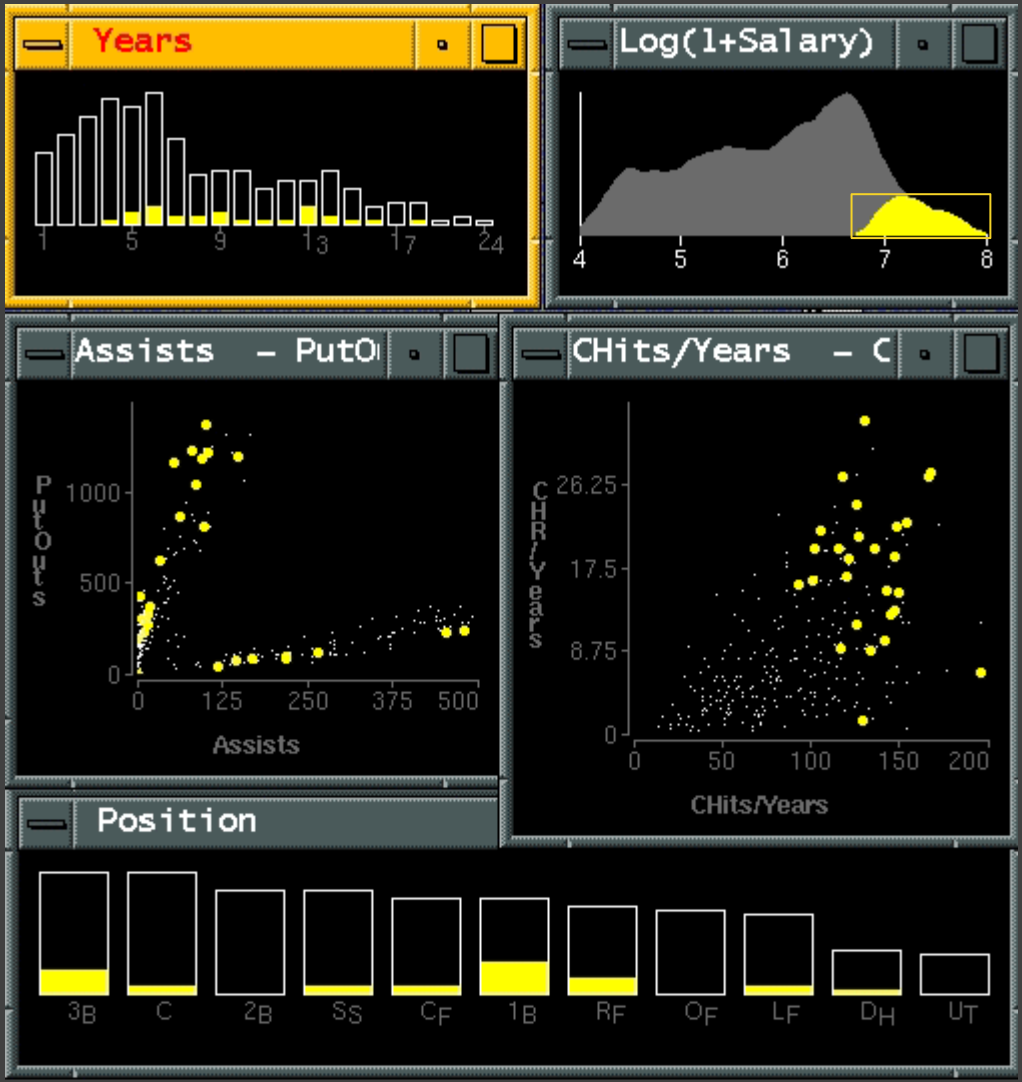
[MacEachren '95, Figure 2.11, p. 38]

Scatterplot Matrix (SPLOM)

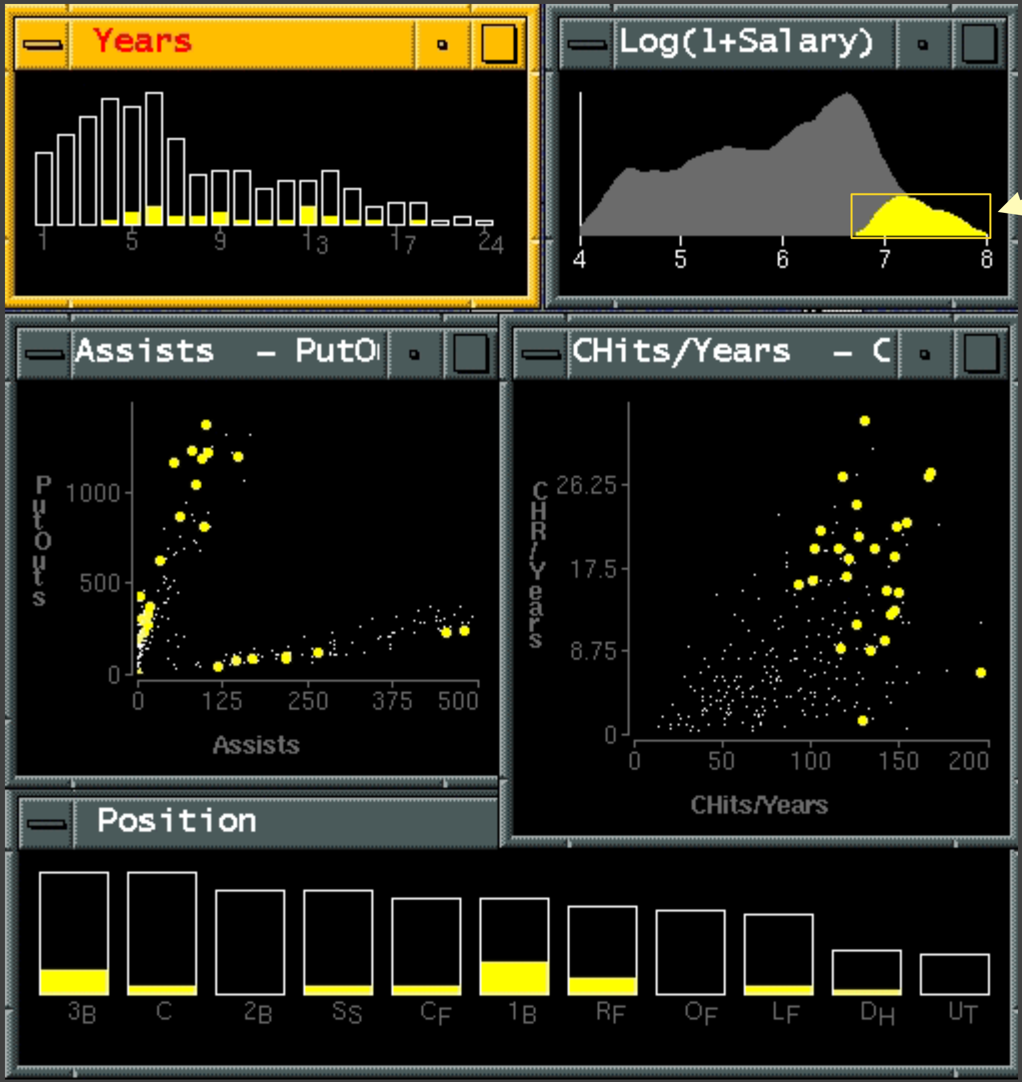


Scatter plots for pairwise comparison of each data dimension.

Multiple Coordinated Views



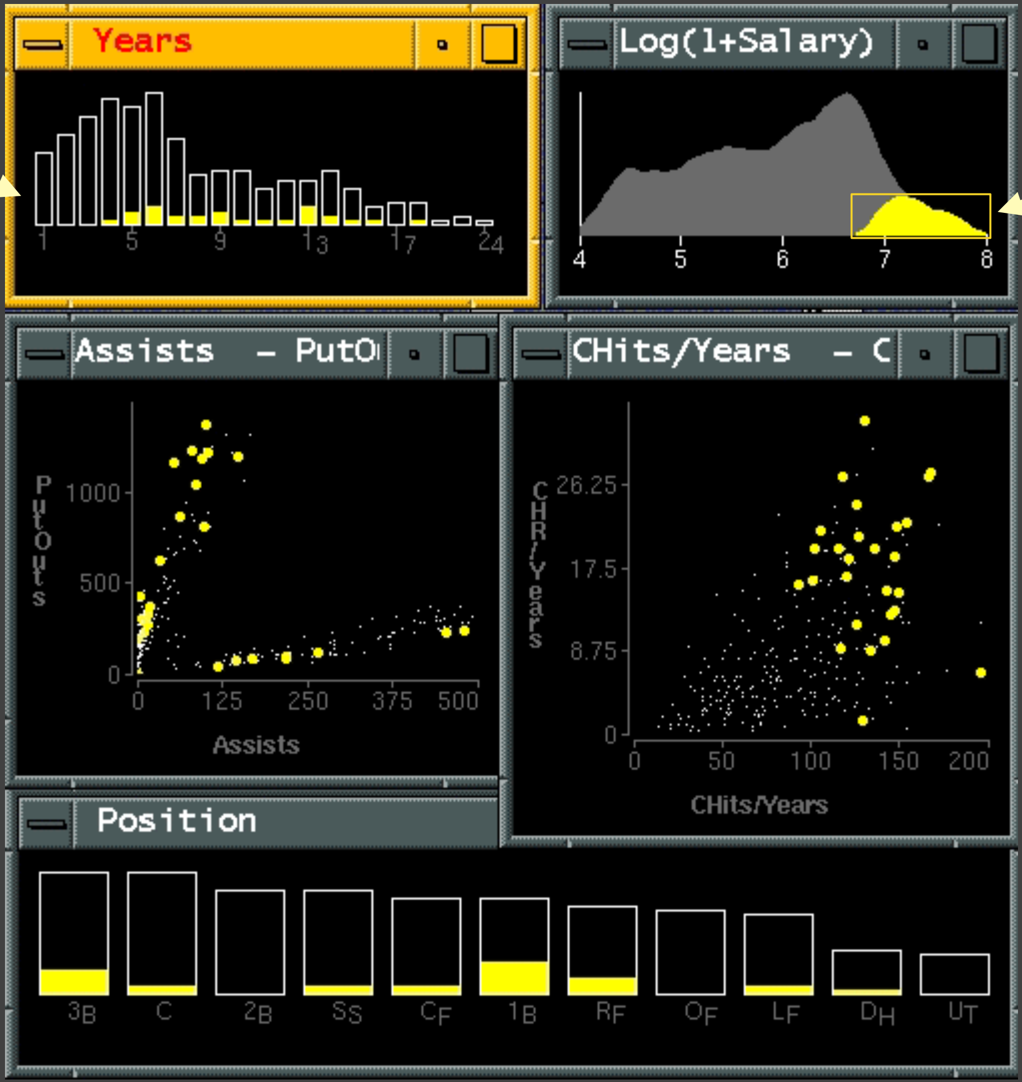
Multiple Coordinated Views



select high salaries

Multiple Coordinated Views

how long in majors



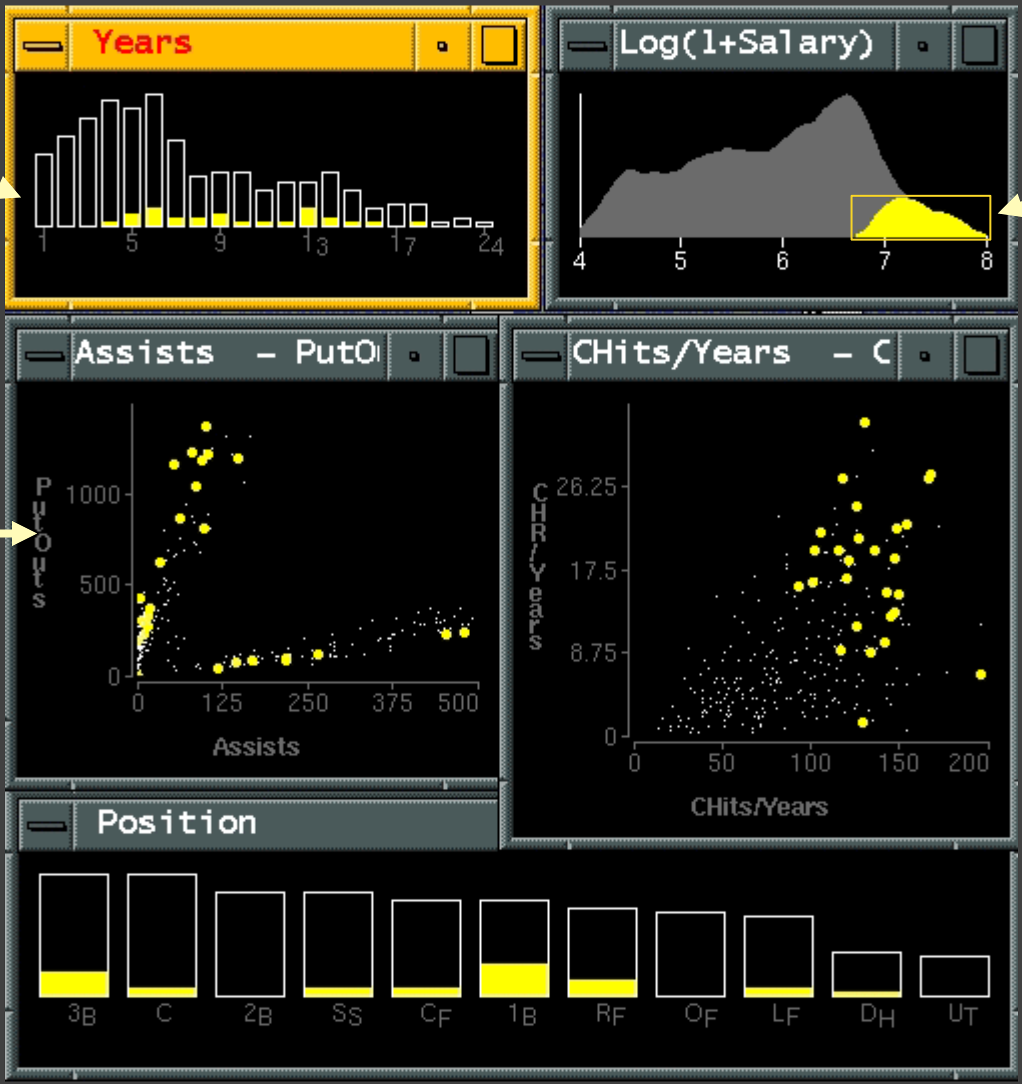
select high salaries

Multiple Coordinated Views

how long in majors

select high salaries

avg assists vs avg putouts (fielding ability)



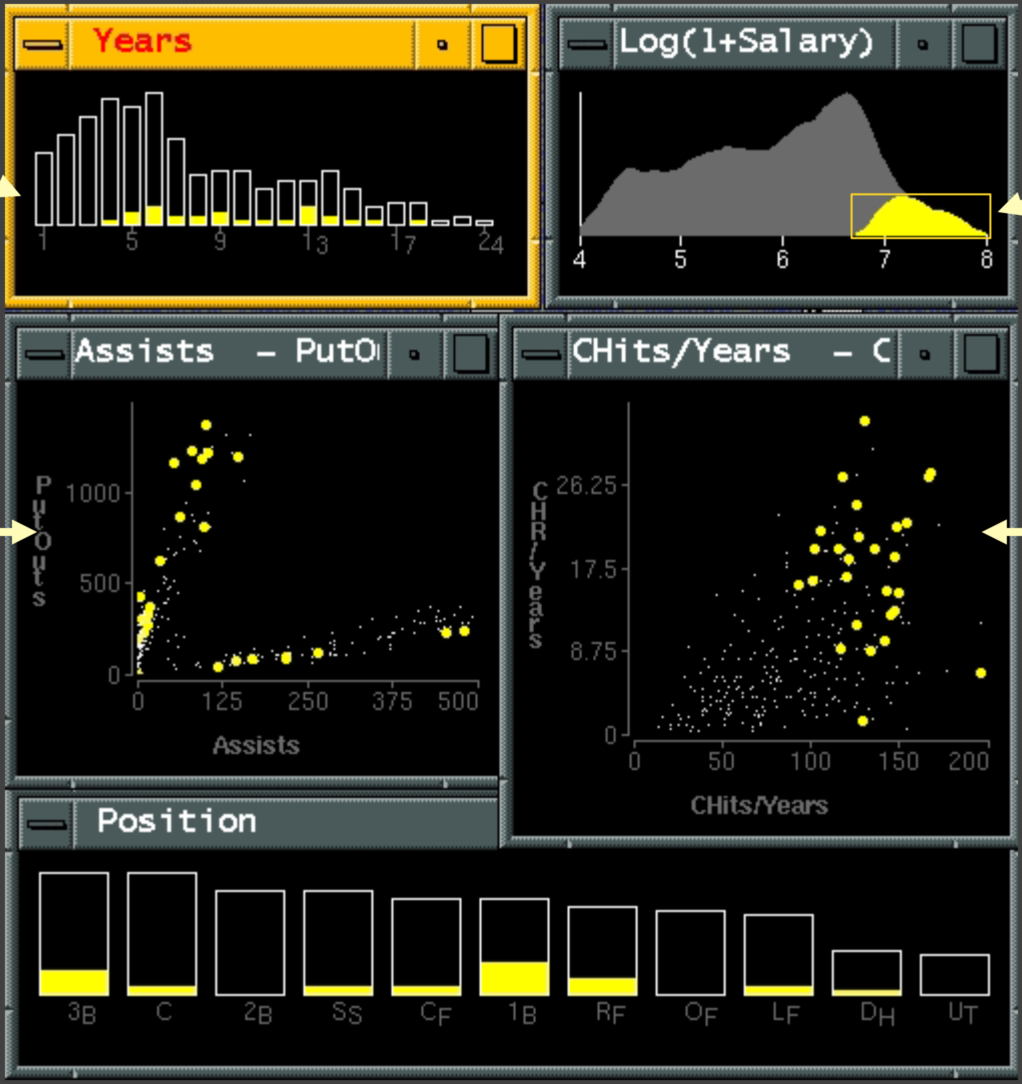
Multiple Coordinated Views

how long in majors

select high salaries

avg assists vs avg putouts (fielding ability)

avg career HRs vs avg career hits (batting ability)



Multiple Coordinated Views

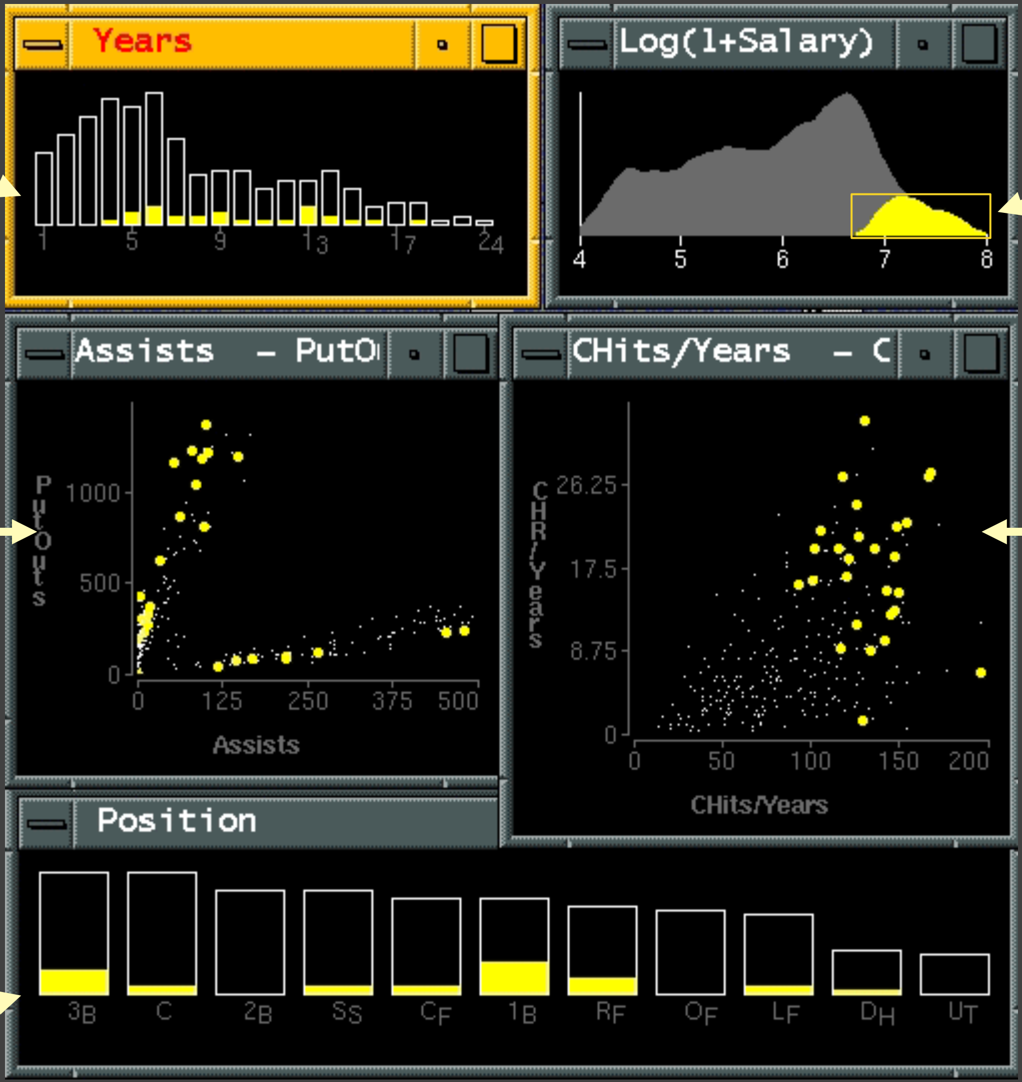
how long in majors

select high salaries

avg assists vs avg putouts (fielding ability)

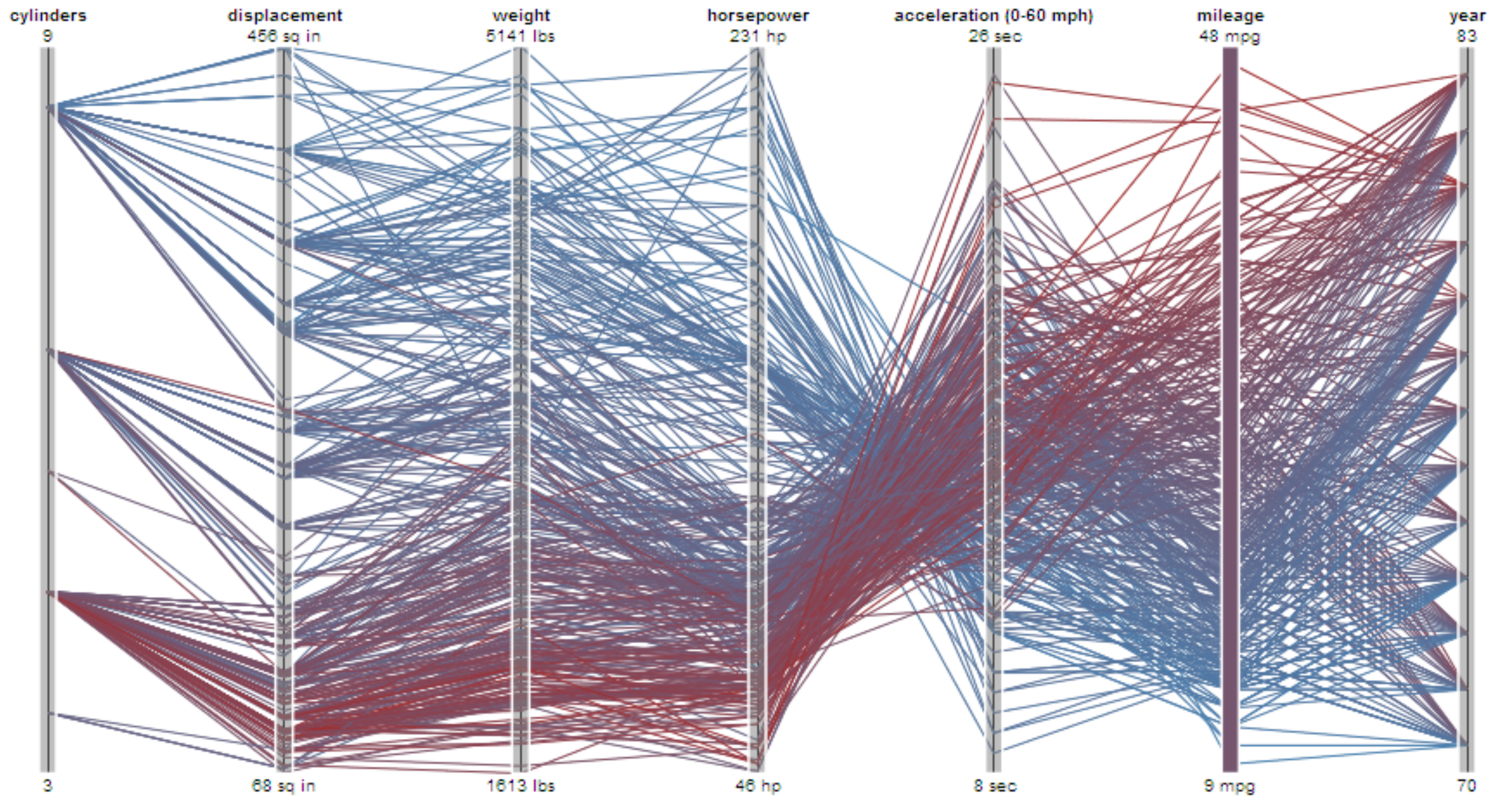
avg career HRs vs avg career hits (batting ability)

distribution of positions played



Parallel Coordinates

Parallel Coordinates [Inselberg]



Parallel Coordinates [Inselberg]

Visualize up to ~two dozen dimensions at once

1. Draw parallel axes for each variable
2. For each tuple, connect points on each axis

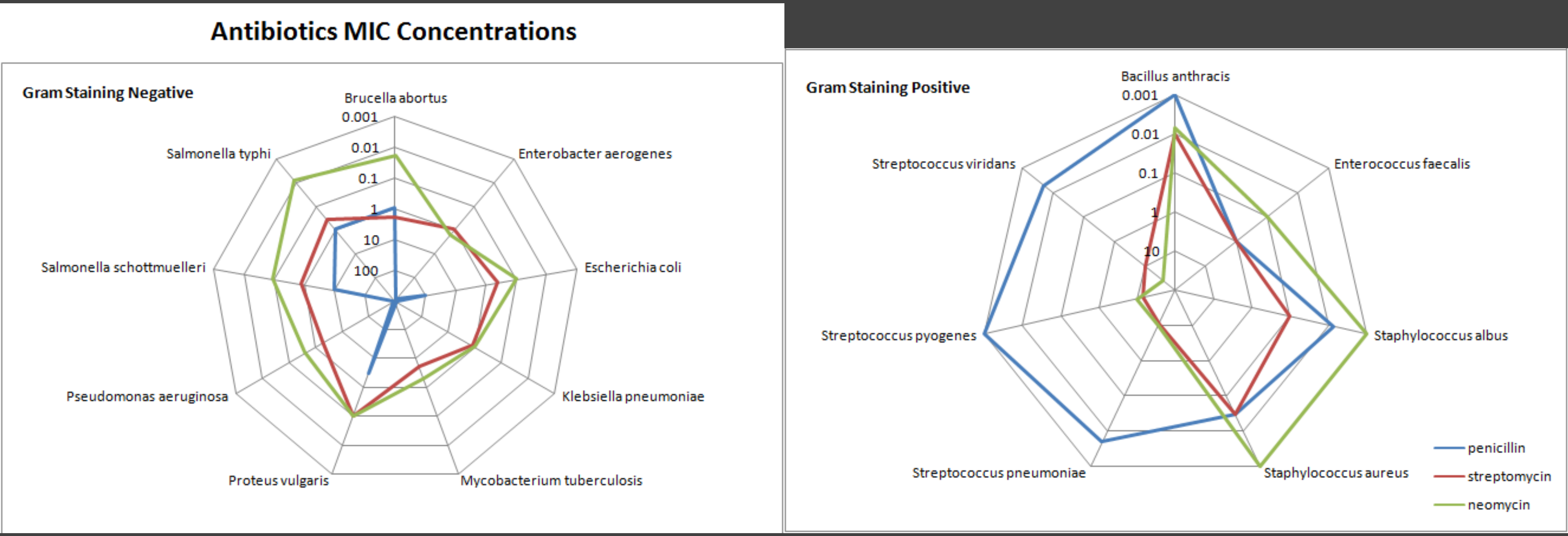
Between adjacent axes: line crossings imply neg. correlation, shared slopes imply pos. correlation.

Full plot can be cluttered. **Interactive selection** can be used to assess multivariate relationships.

Highly sensitive to axis **scale** and **ordering**.

Expertise required to use effectively!

Radar Plot / Star Graph

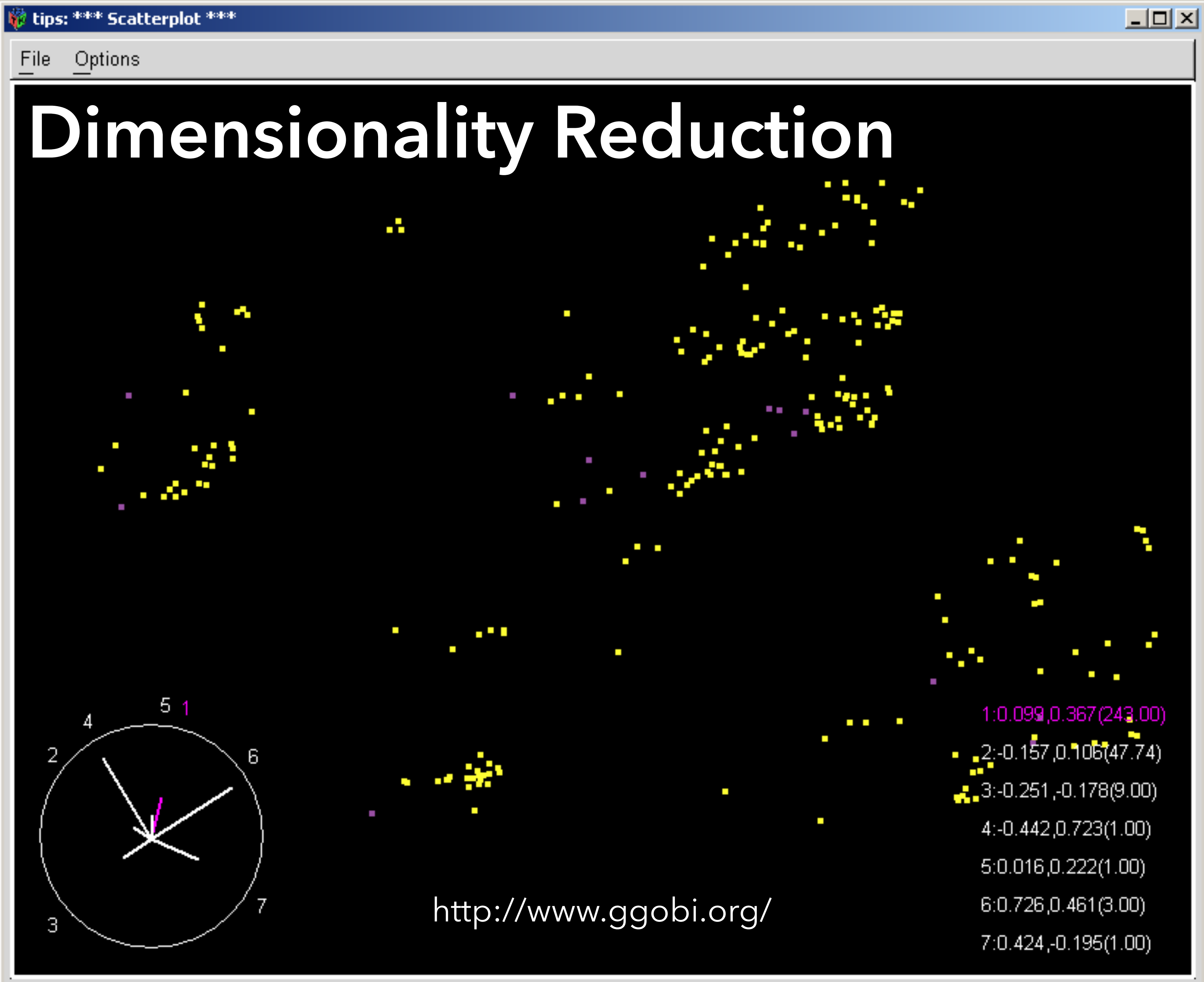


"Parallel" dimensions in polar coordinate space

Best if same units apply to each axis

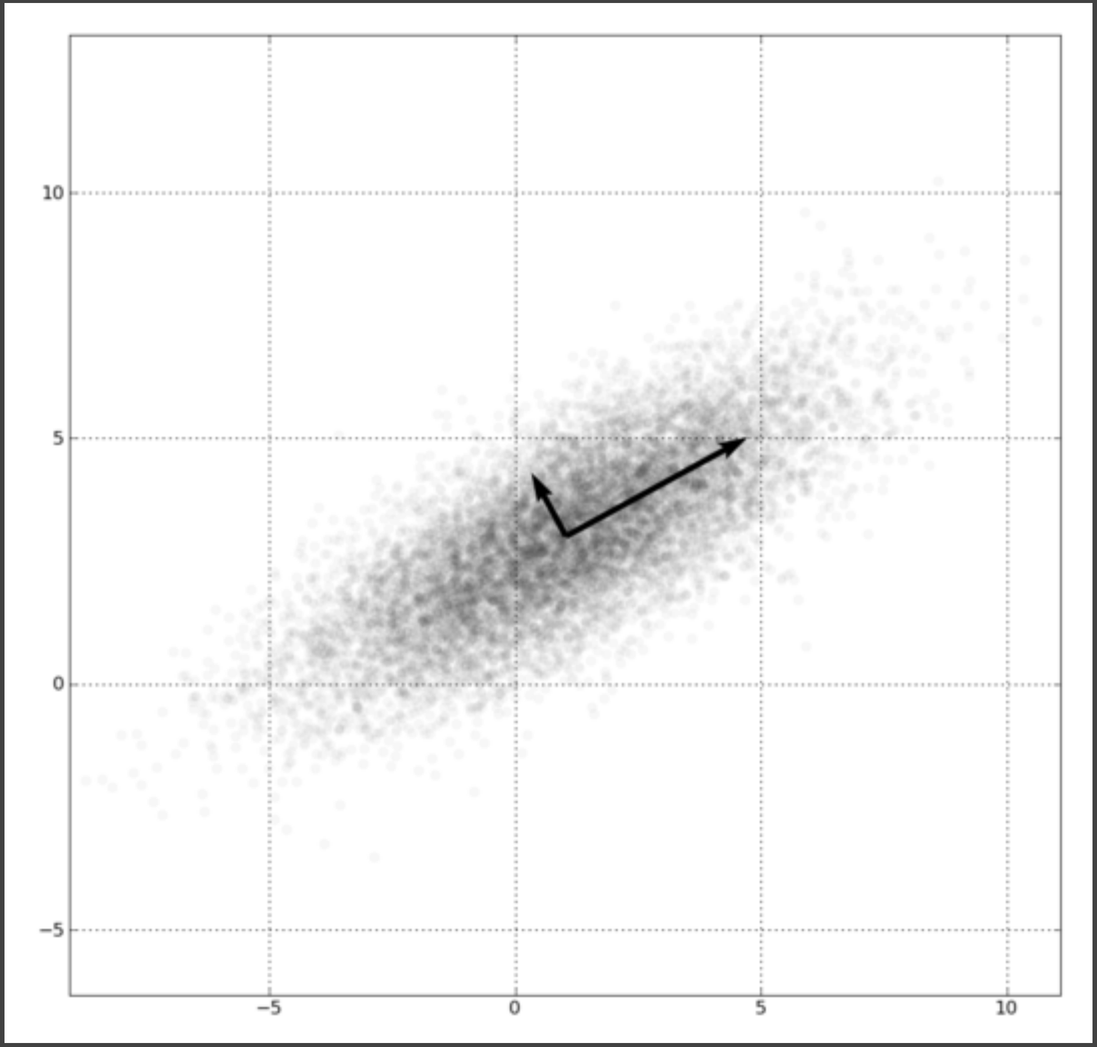
Dimensionality Reduction

Dimensionality Reduction



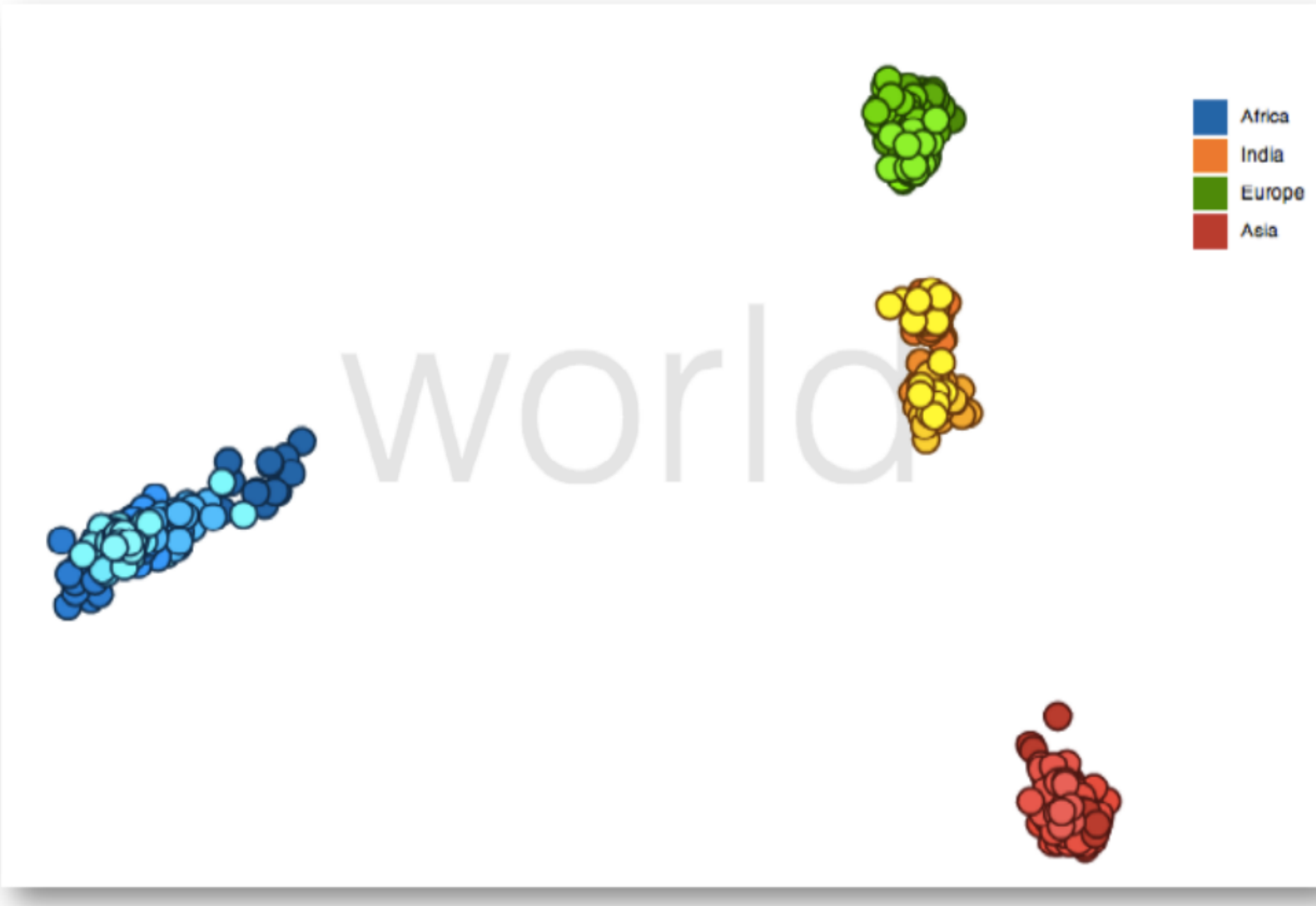
http://www.ggobi.org/

Principal Components Analysis



1. Mean-center the data.
2. Find ⊥ basis vectors that maximize the data variance.
3. Plot the data using the top vectors.

PCA of Genomes [Demiralp et al. '13]



Many Reduction Techniques!

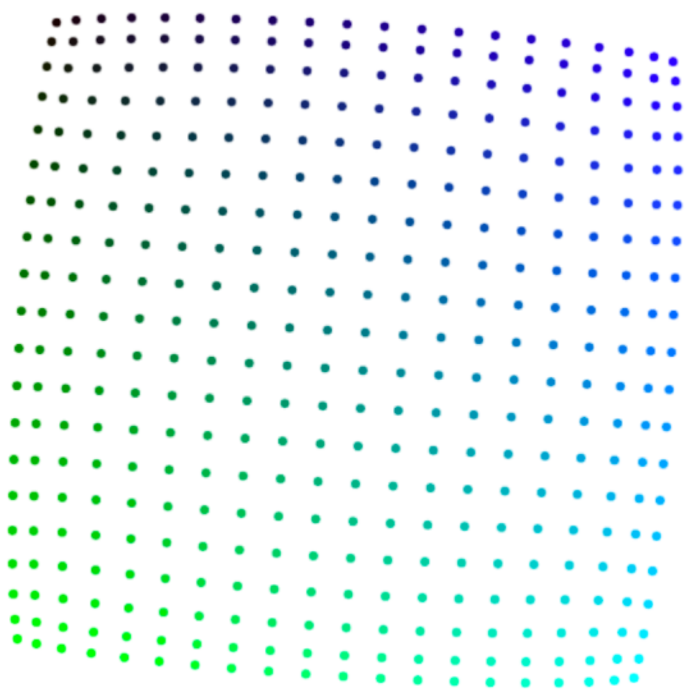
General Strategies:
Matrix Factorization
Nearest Neighbor (Topological) Methods

Popular Techniques:
Principal Components Analysis (PCA)
t-Dist. Stochastic Neighbor Embedding (t-SNE)
Uniform Manifold Approx. & Projection (UMAP)

Distill ABOUT PRIZE SUBMIT

How to Use t-SNE Effectively

Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.



Step 1,910

A square grid with equal spacing between points. Try convergence at different sizes.

Points Per Side 20

Perplexity 10

Epsilon 5

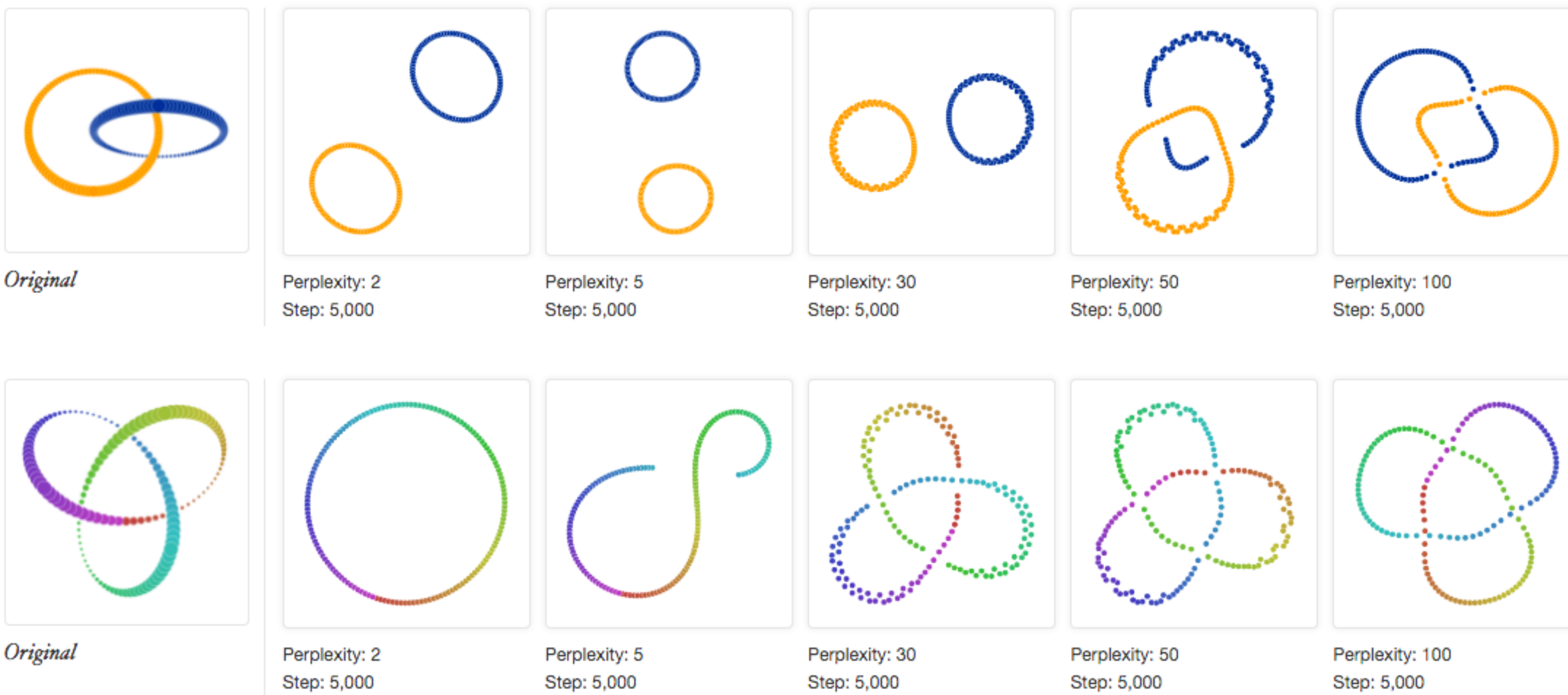
MARTIN WATTENBERG Google Brain
FERNANDA VIÉGAS Google Brain
IAN JOHNSON Google Cloud

Oct. 13 2016

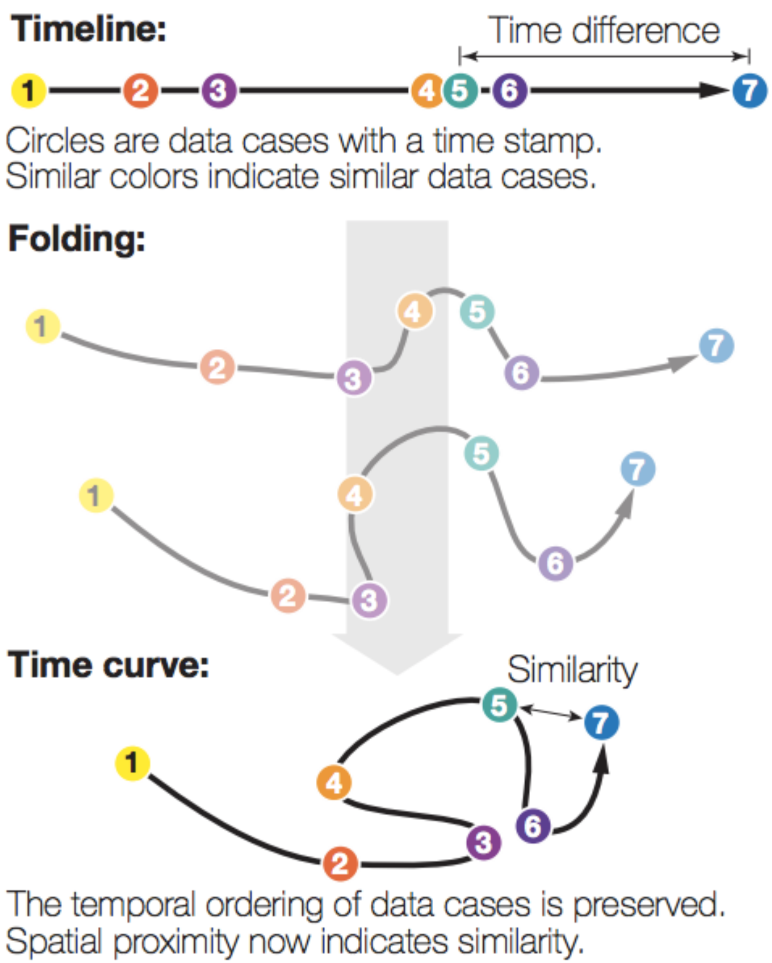
Citation: Wattenberg, et al., 2016

distill.pub

Visualizing t-SNE [Wattenberg et al. '16]

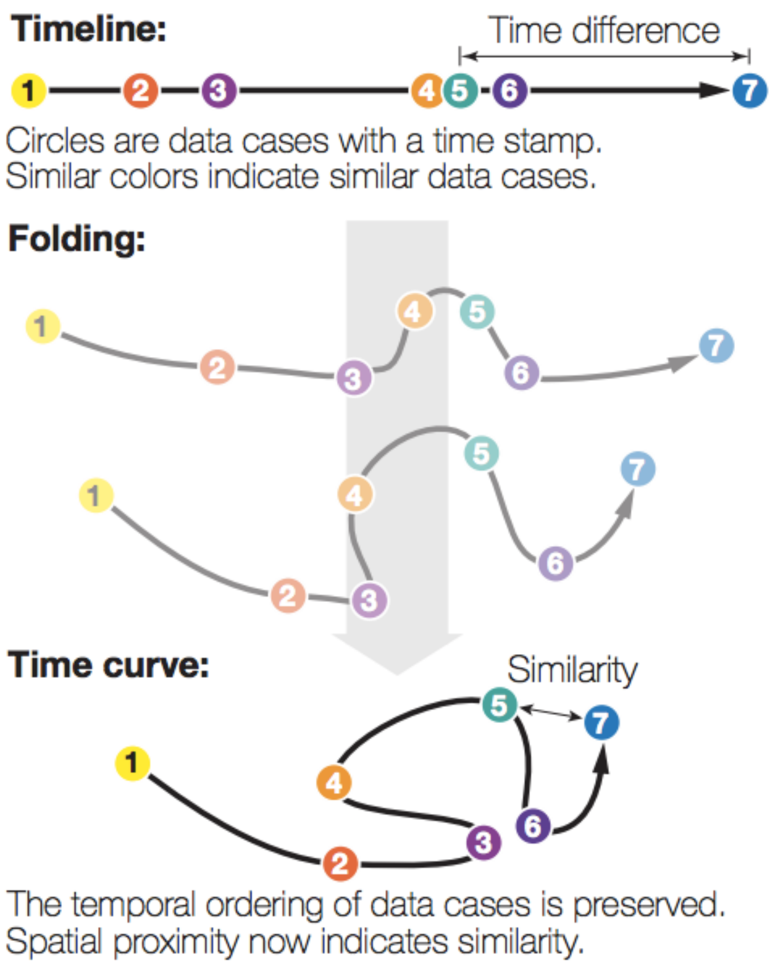


Time Curves [Bach et al. '16]

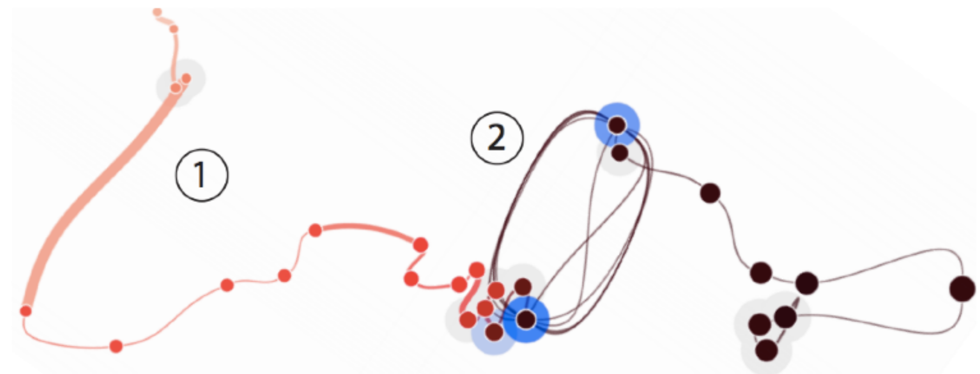


(a) Folding time

Time Curves [Bach et al. '16]

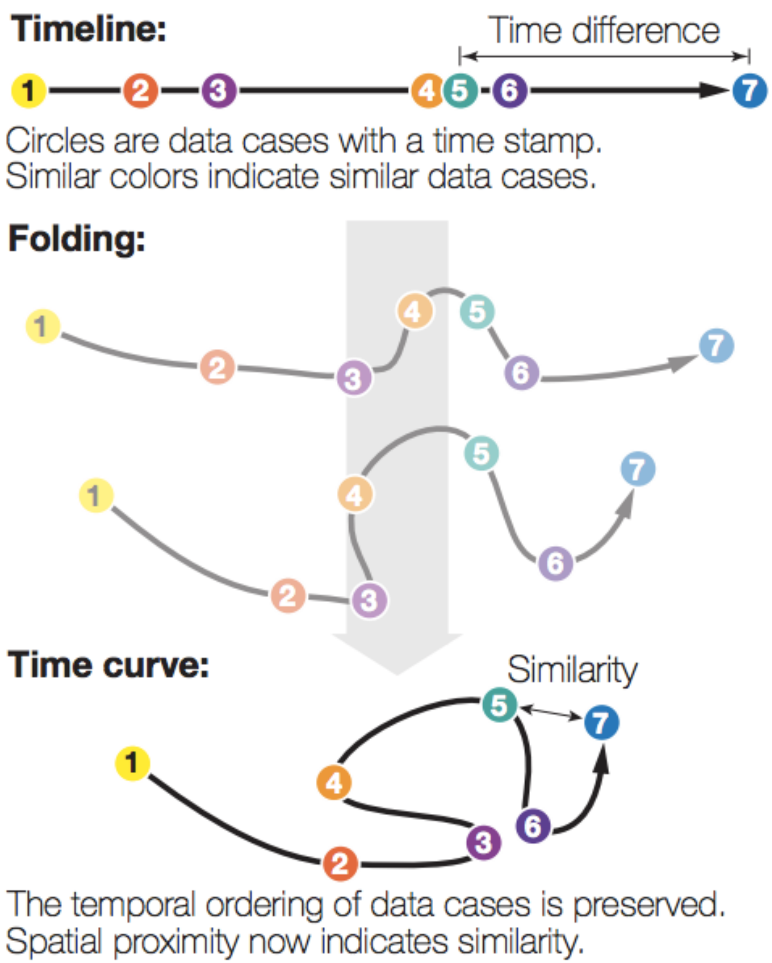


(a) Folding time

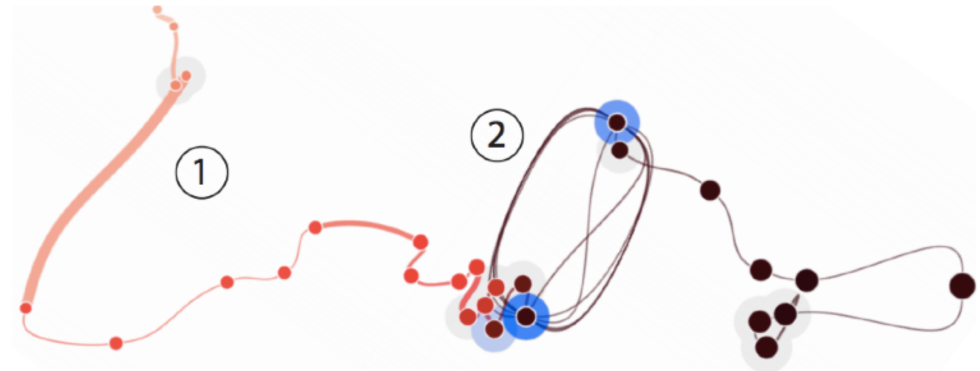


Wikipedia "Chocolate" Article

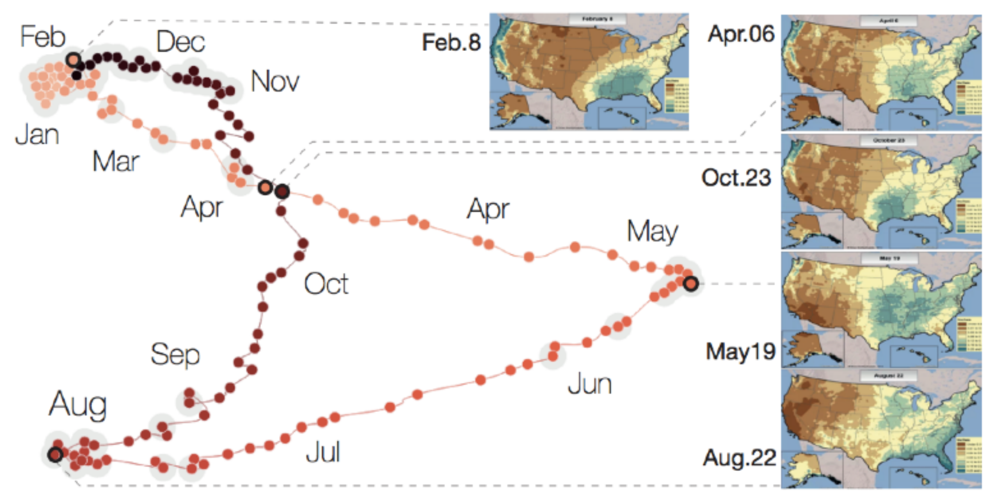
Time Curves [Bach et al. '16]



(a) Folding time



Wikipedia “Chocolate” Article



U.S. Precipitation over 1 Year

Visual Encoding Design

Use **expressive** and **effective** encodings
Avoid **over-encoding**
Reduce the problem space
Use **space** and **small multiples** intelligently
Use **interaction** to generate *relevant* views

Rarely does a single visualization answer all questions. Instead, the ability to generate appropriate visualizations quickly is critical!