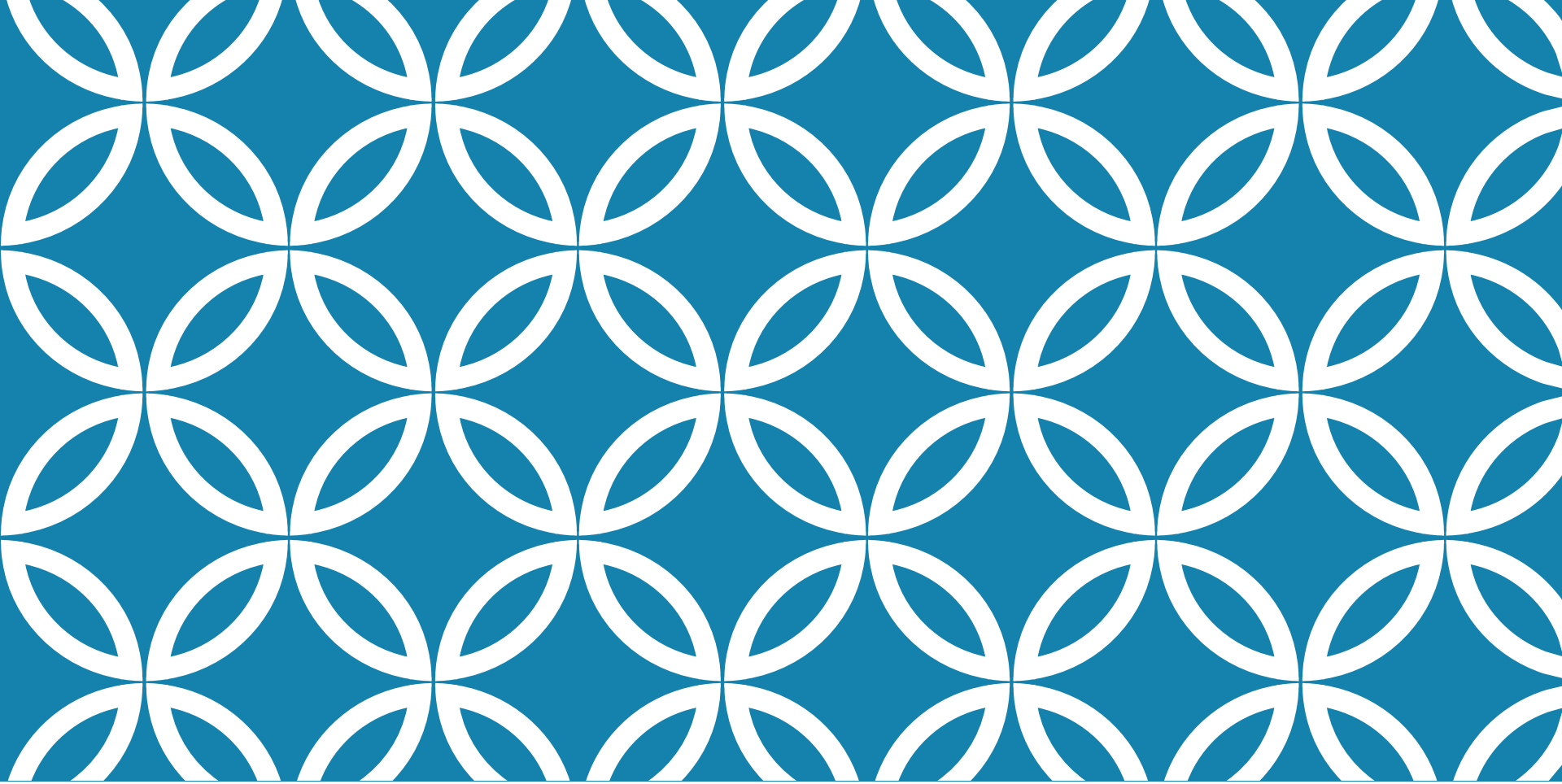


VISUALIZATION IN R WITH GGPLOT2

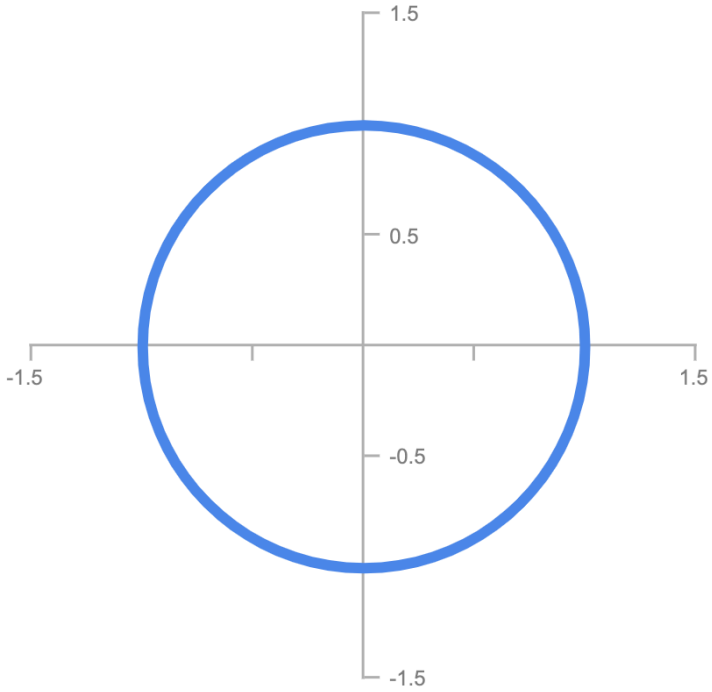
Agnes Chang

PSYC GR6130, Fall 2019

AGENDA

- Why visualize? Exploratory vs. explanatory goals.
- Data models: categorical, ordinal, quantitative
- Visual encoding, and what's effective?
- Grammar of graphics

WHY VISUALIZE?



$$x^2 + y^2 = 1$$

Via Shan Carter, openvisconf

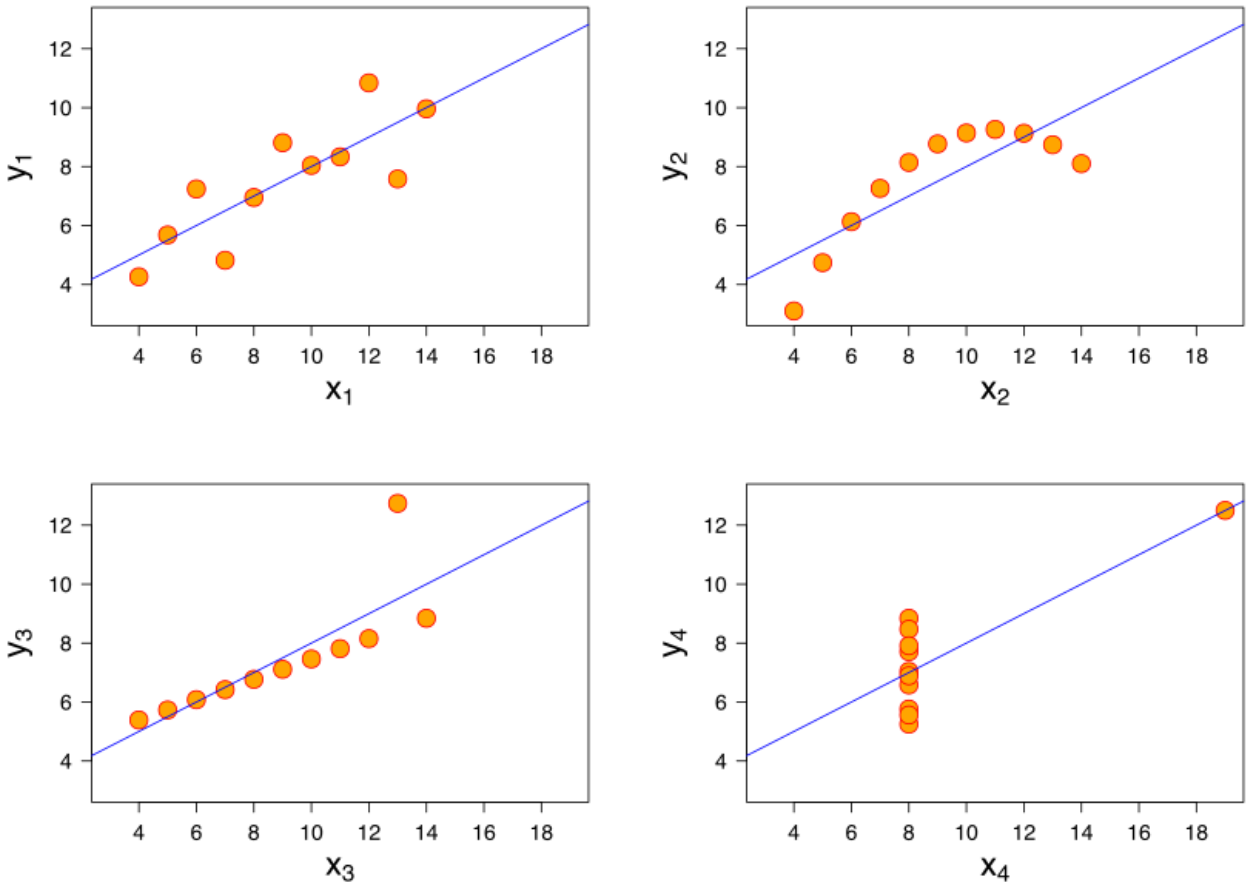
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08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

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

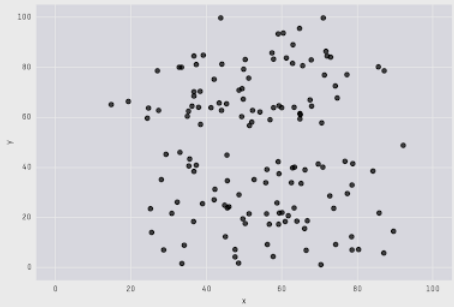
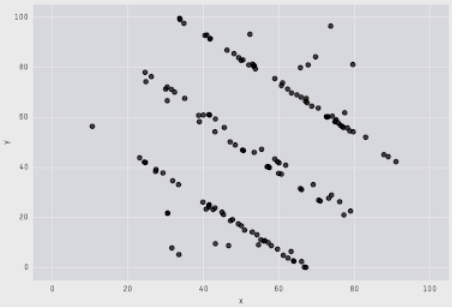
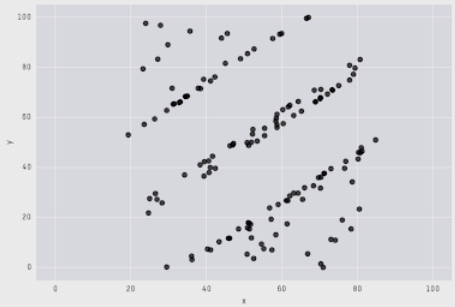
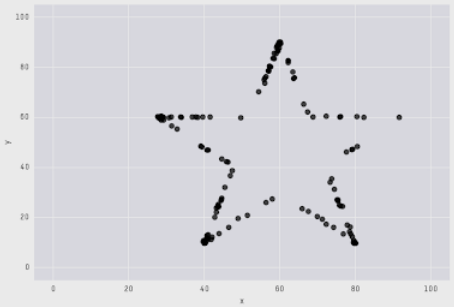
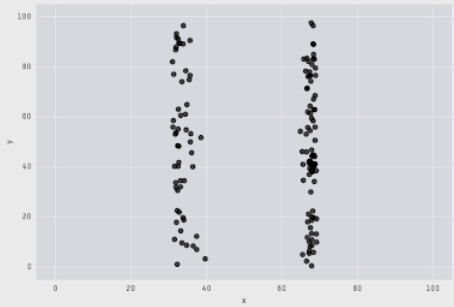
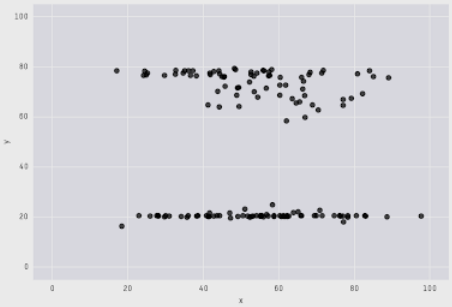
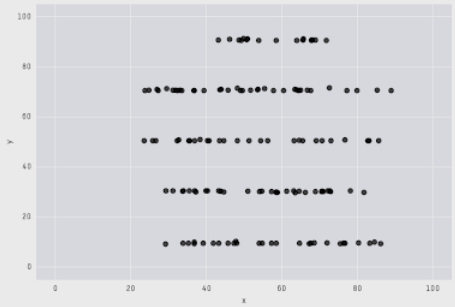
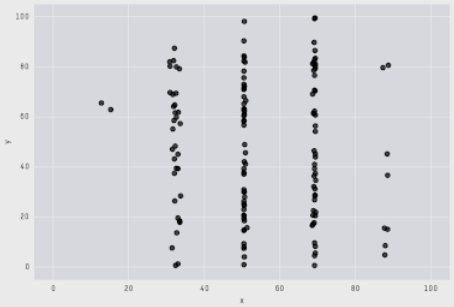
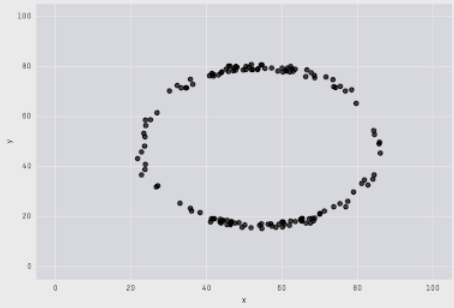
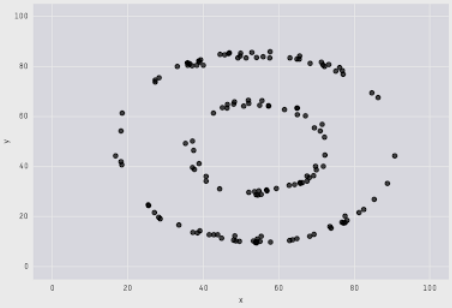
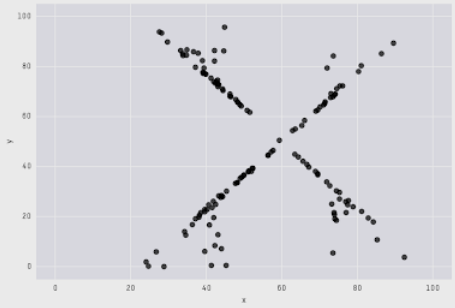
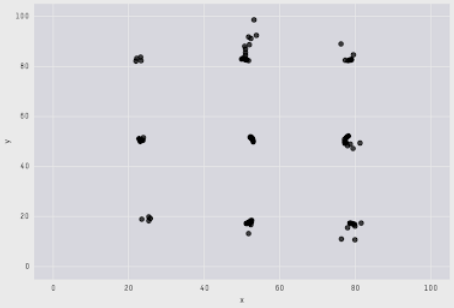
Mean & Variance
uX = 9.0, σX = 11
uY = 7.5, σY = 4.125

Linear Regression
Y = 3 + 0.5X
R^2 = 0.67

ANSCOMBE'S QUARTET



X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

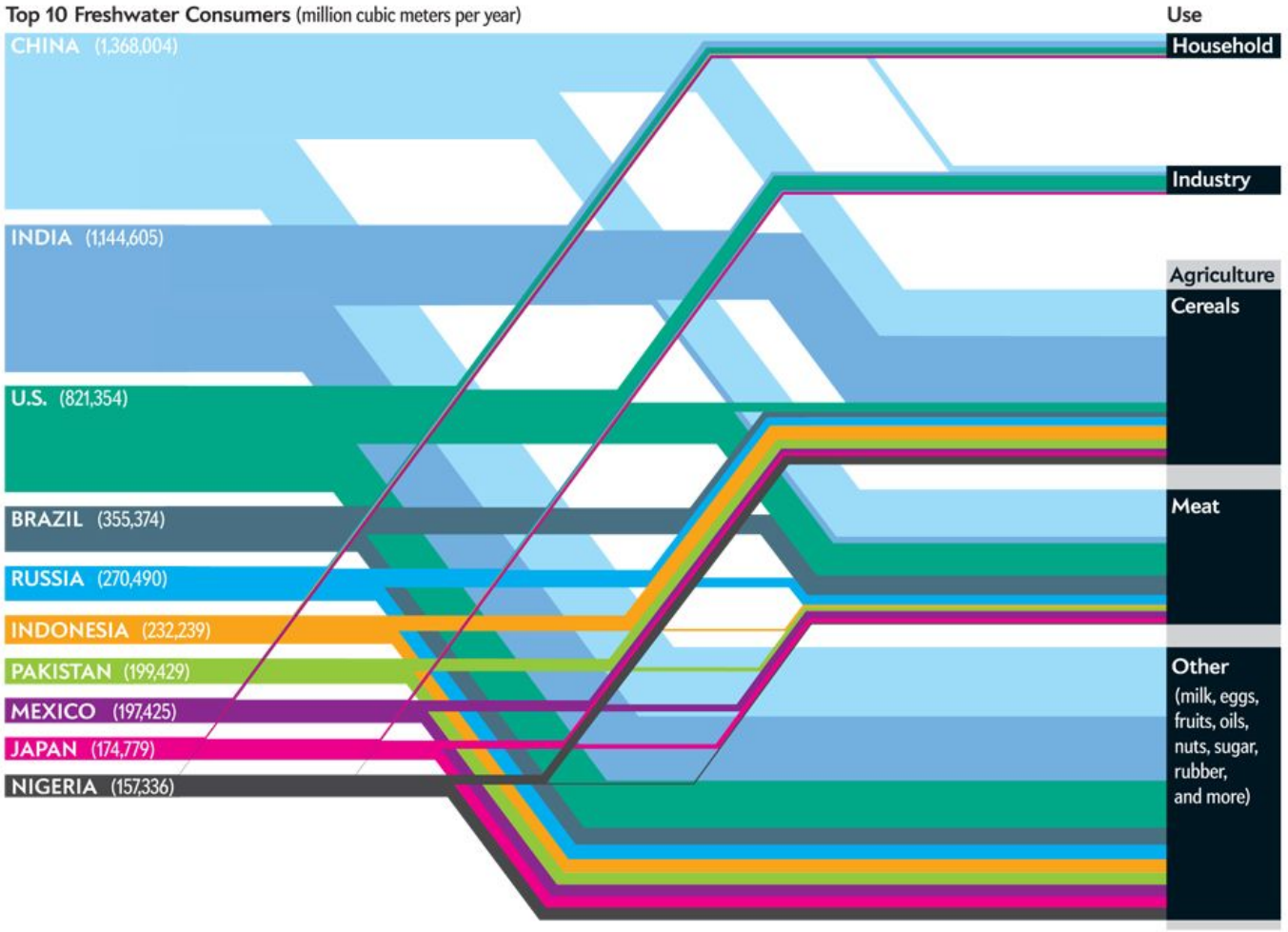













https://www.autodeskresearch.com/publications/samestats

EXPLORATORY: “SPLOM”

https://observablehq.com/@d3/brushable-scatterplot-matrix

EXPLANATORY: SCI. AM.



https://www.scientificamerican.com/article/water-in-water-out/

GOALS OF VISUALIZATION

Exploratory

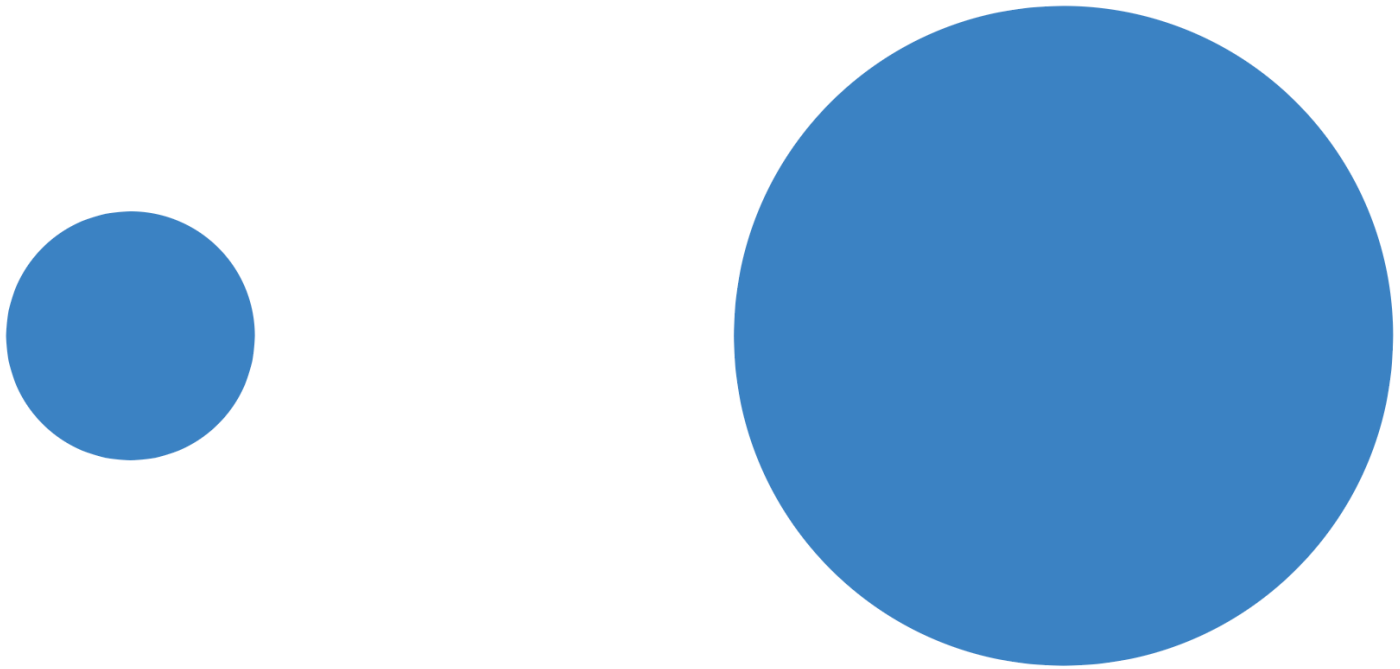
- to uncover a relationship in the data
- to analyze data

Explanatory

- to communicate a relationship in the data
- to present data

BUT THERE'S AN ART AND A SCIENCE TO IT….

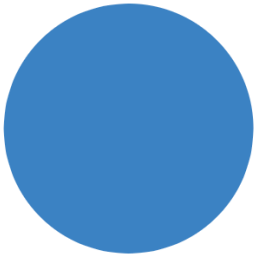
COMPARE AREA OF CIRCLES

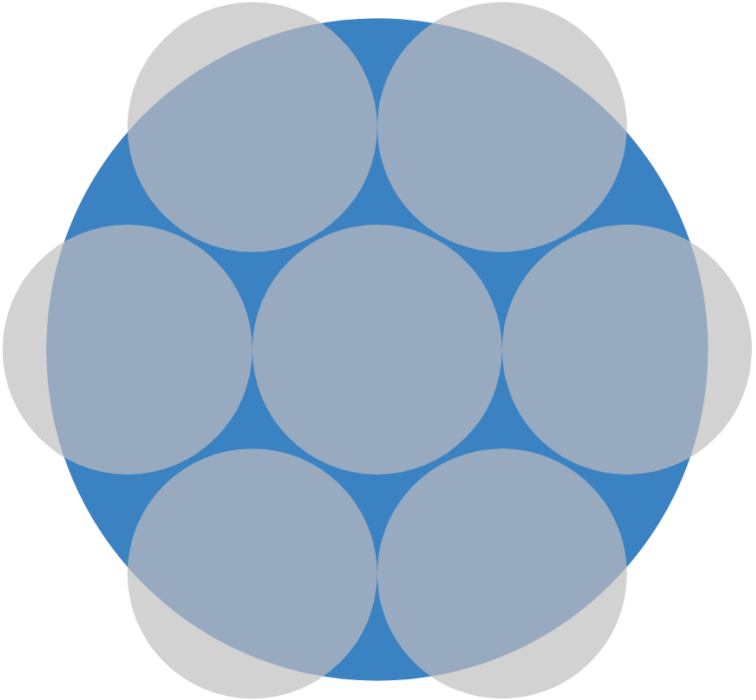


Via Jeff Heer

COMPARE LENGTH OF BARS

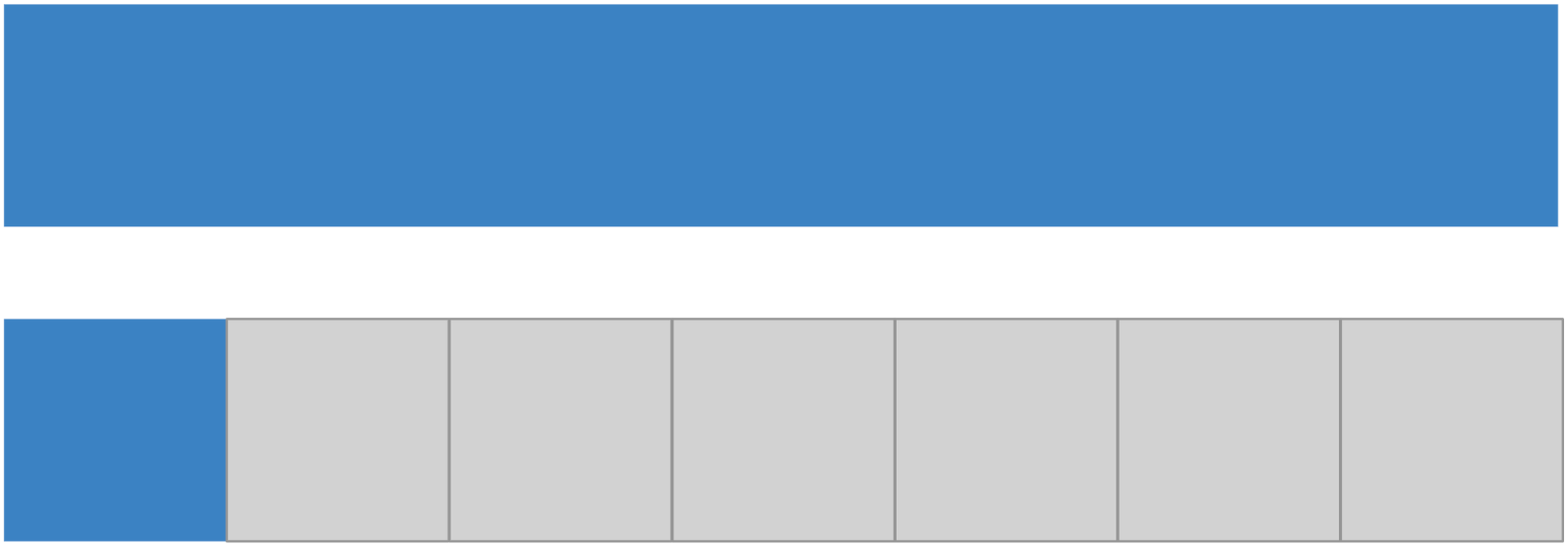
COMPARE AREA OF CIRCLES



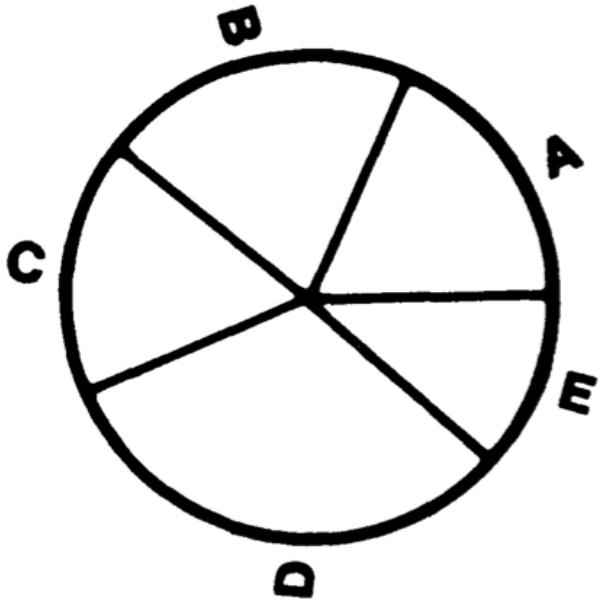


Via Jeff Heer

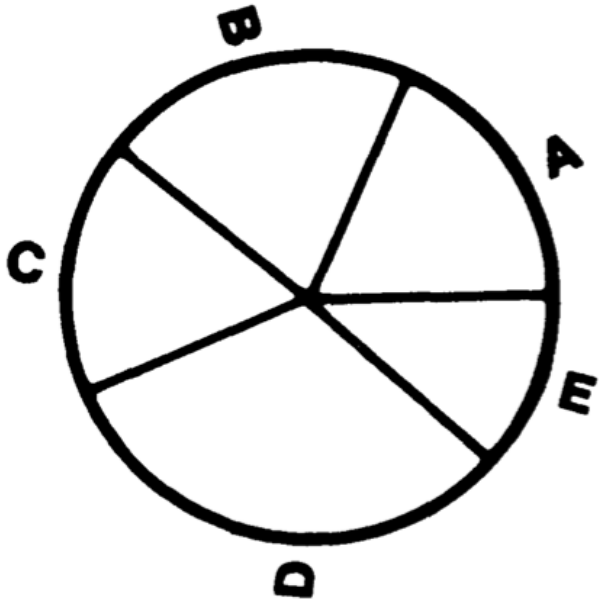
COMPARE LENGTH OF BARS

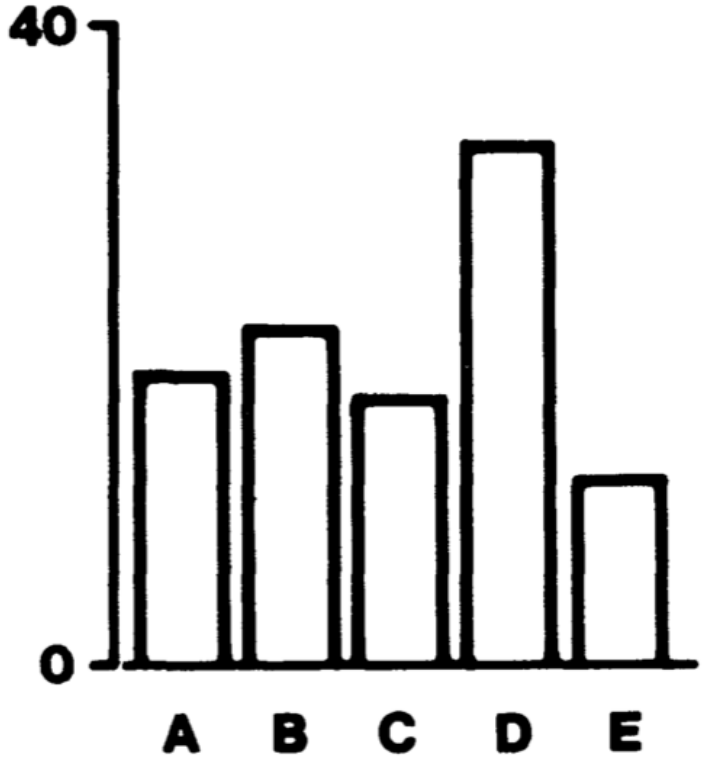


WHICH IS LARGER, A OR C?



WHICH IS LARGER, A OR C?





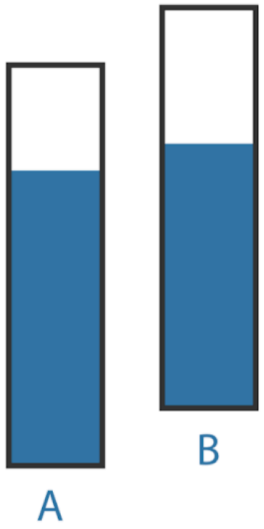
MCGILL CLEVELAND RANKING ACCURACY

Unframed
Unaligned

Munzner. *Visualization Analysis and Design* (2015)

MCGILL CLEVELAND RANKING ACCURACY

Unframed
Unaligned

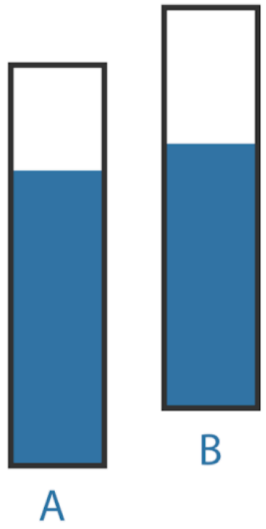


Framed
Unaligned

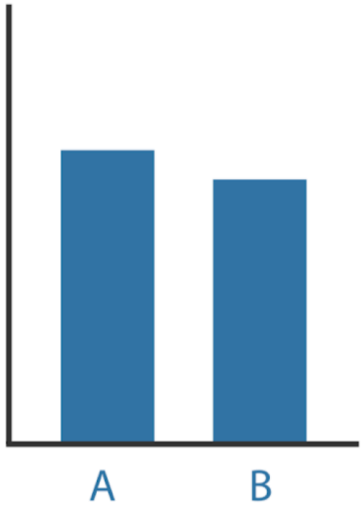
Munzner. *Visualization Analysis and Design* (2015)

MCGILL CLEVELAND RANKING ACCURACY

Unframed
Unaligned



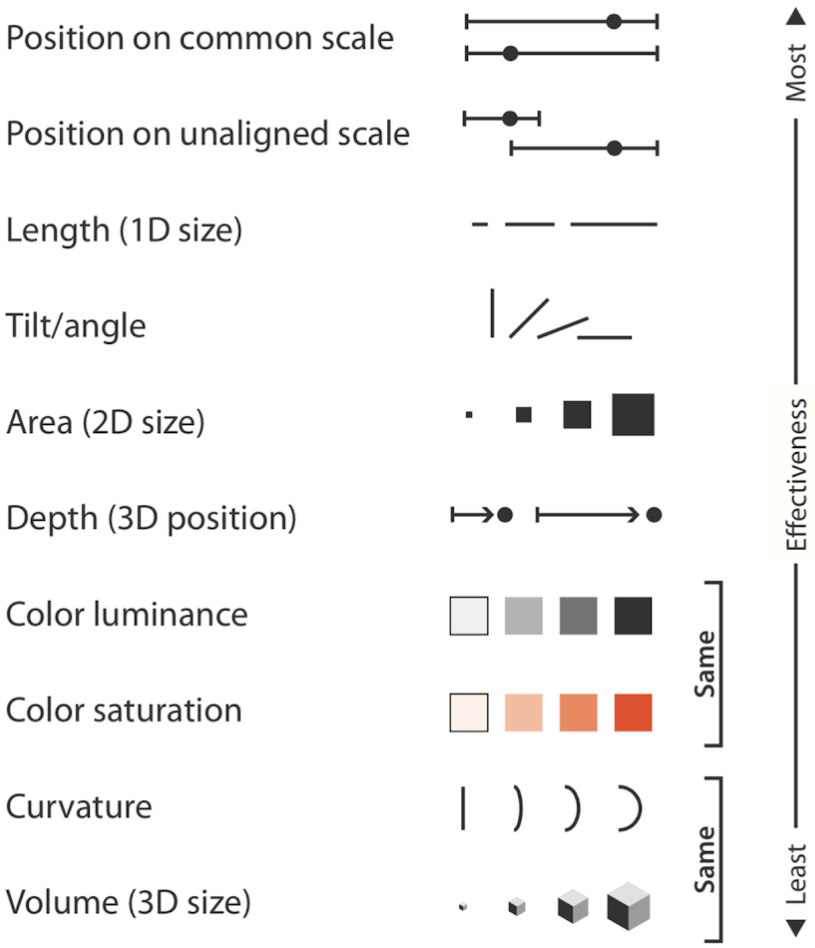
Framed
Unaligned



Framed
Aligned

Munzner. *Visualization Analysis and Design* (2015)

PERCEPTION RANKING



Munzner. *Visualization Analysis and Design* (2015)

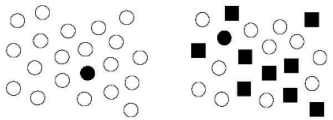
WHAT IS DATA VIZ?

- Transform data into visual encodings.
- Using special properties of the visual system* to help us think.

**Corollary: All visualizations are made from a series of compromises.*

Via Wattenberg & Viegas

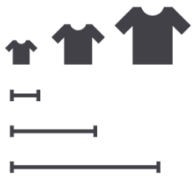
Perceptual Properties



VISUAL ENCODING

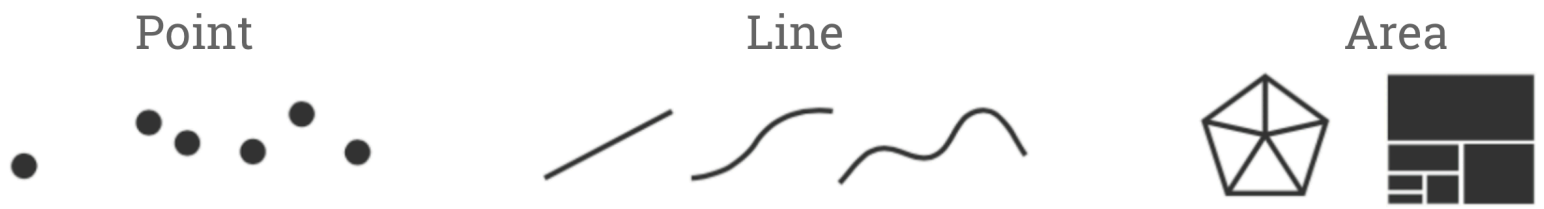
Marks & Channels

Data Types



MARKS

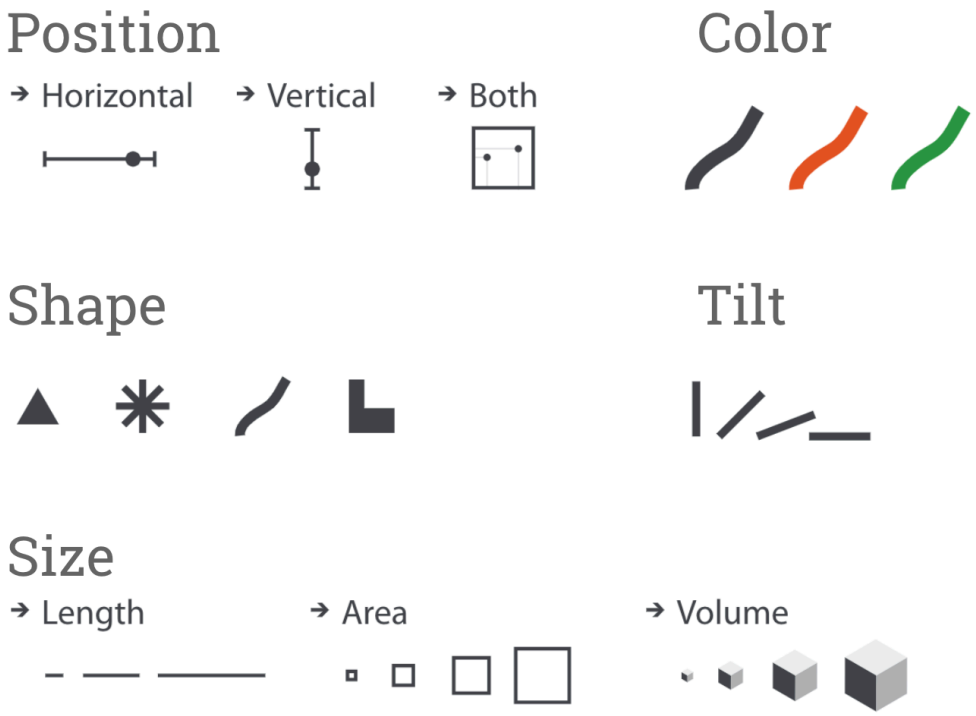
= Basic graphical element in image
(a.k.a. geom in ggplot)



Graphics from Munzner. *Visualization Analysis and Design* (2015)

CHANNELS

= Ways to control appearance of marks
(a.k.a. aes in ggplot)



DATA TYPES

= Classifies the semantic meaning of the data

Categorical

Fruit (apple, pear, kiwi…)
Cities (NYC, SF, LA…)

Ordinal

Months (Jan, Feb, Mar…)
Sizes (S, M, L, XL…)

Quantitative

Lengths (1”, 2.5”, 5.14”…)
Population

DATA TYPES EXAMPLE

Data Set: NYC Daily Temperatures, 2017

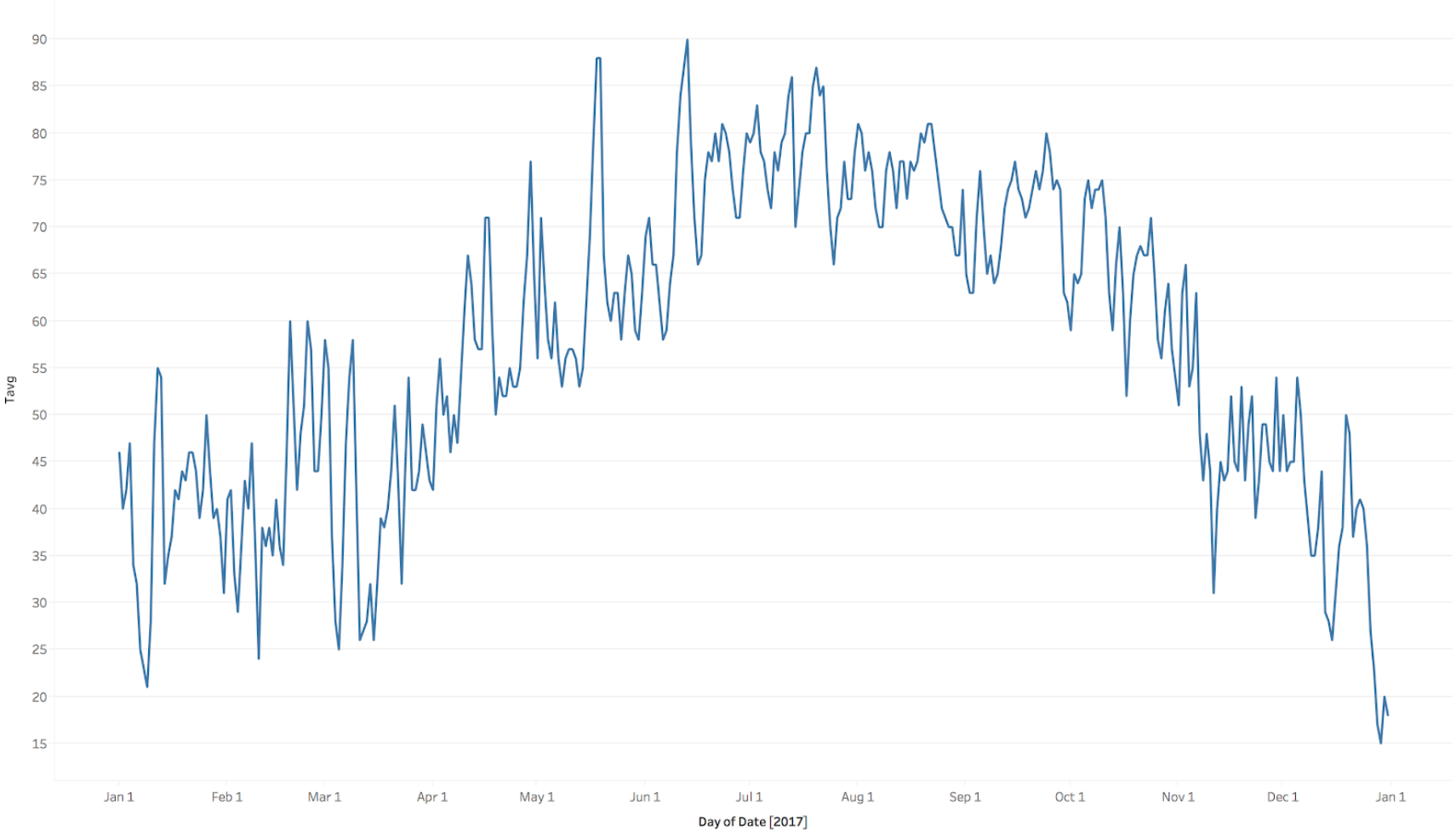
Data Model: 21, 28, 47, 55, … (integers)

Conceptual Model: Temperature (°F)

Data Type:

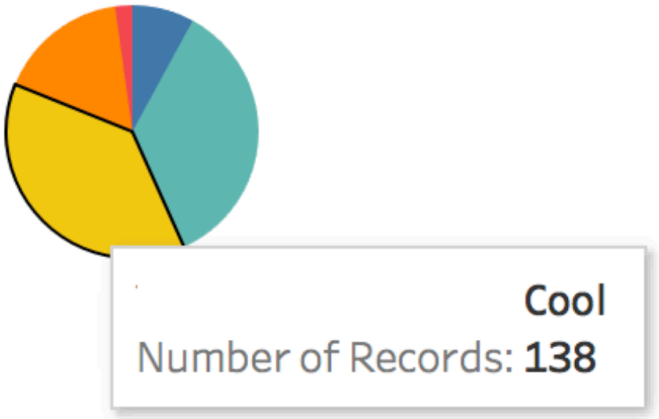
- Freezing vs. Not-Freezing (C)
- Hot, Warm, Cold, Freezing (O)
- Temperature Value (Q)

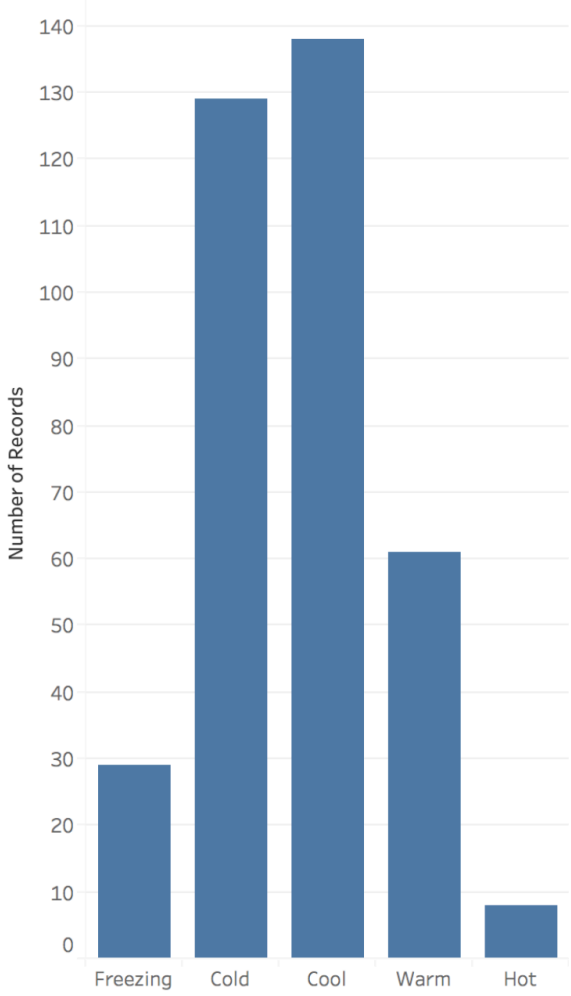
E.G. QUANT X QUANT



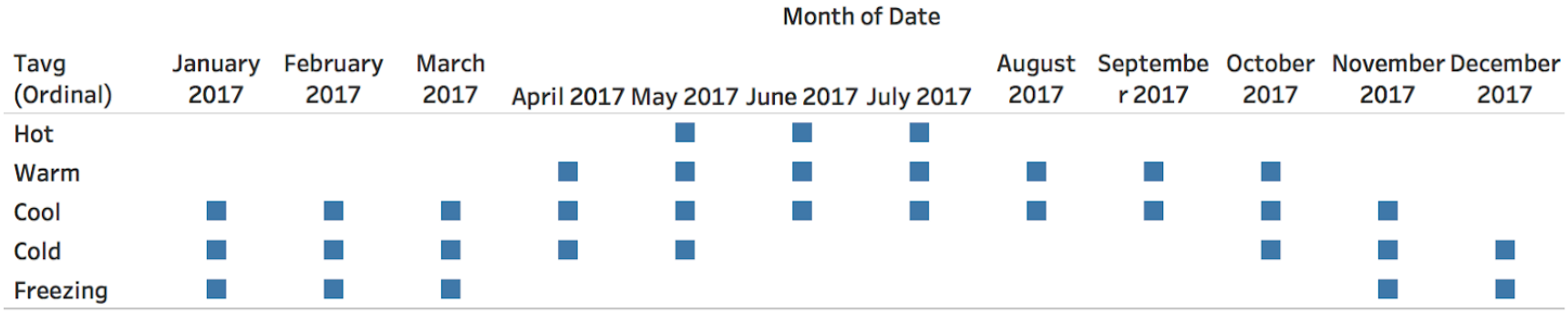
E.G. ORDINAL X QUANT

Freezing	<32
Cold	32–54
Cool	55–74
Warm	75–85
Hot	85+

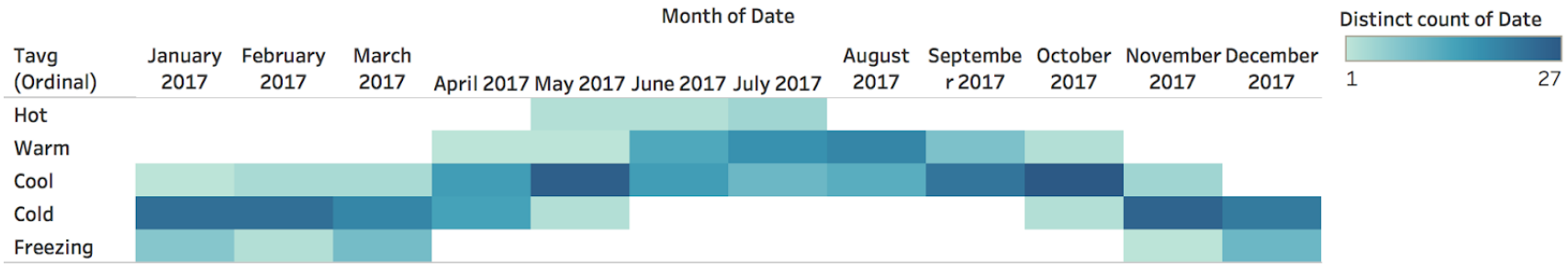




E.G. ORDINAL X ORDINAL



E.G. ORD X ORD X QUANT

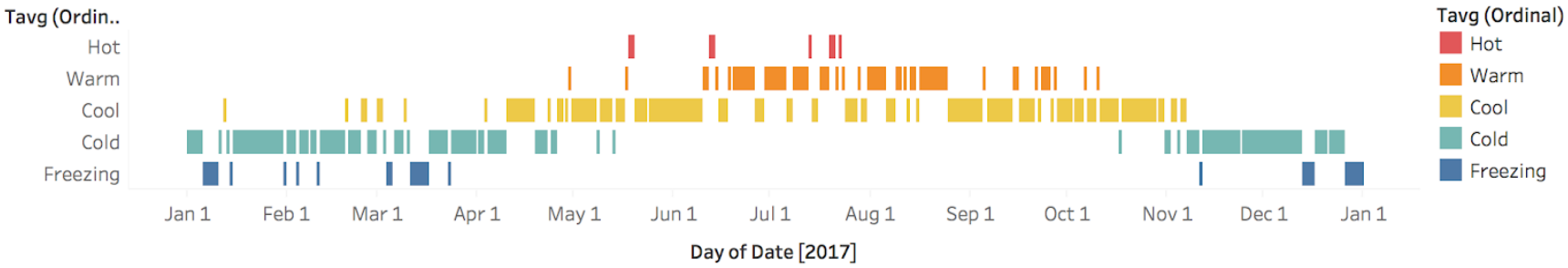


X-axis: month (O)

Y-axis: temperature (O)

Color: count of days (Q)

E.G. ORD X QUANT

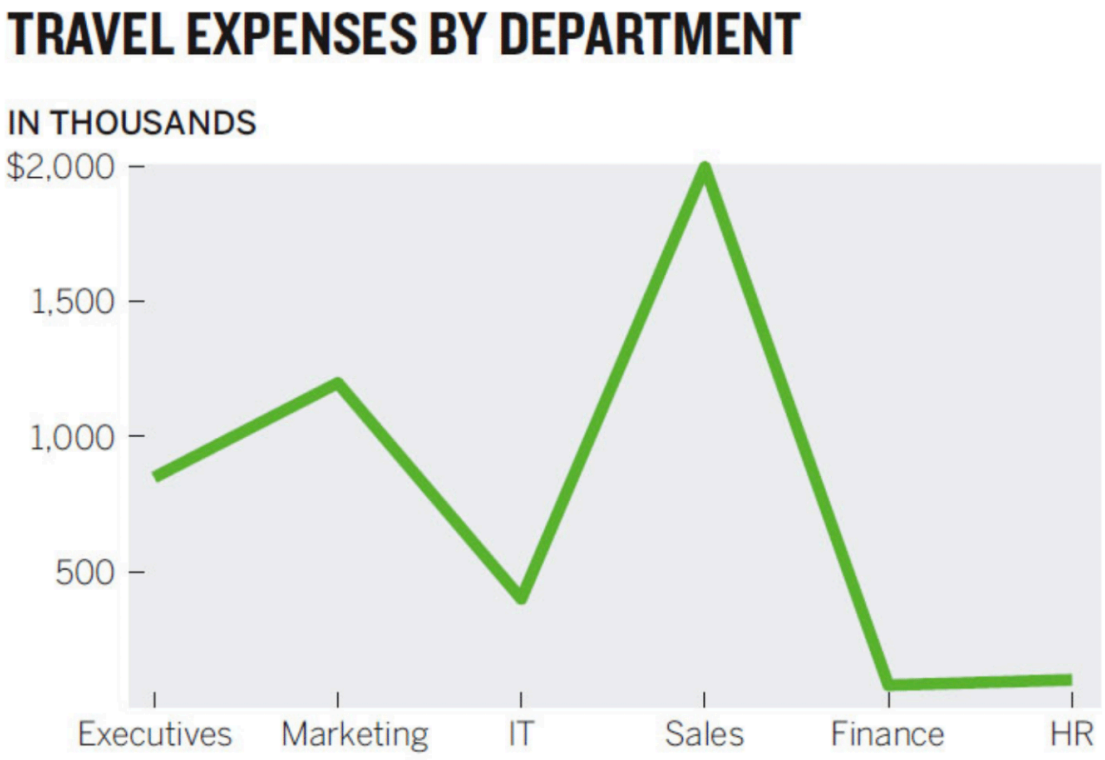


X-axis: date (Q)

Y-axis: temperature (O)

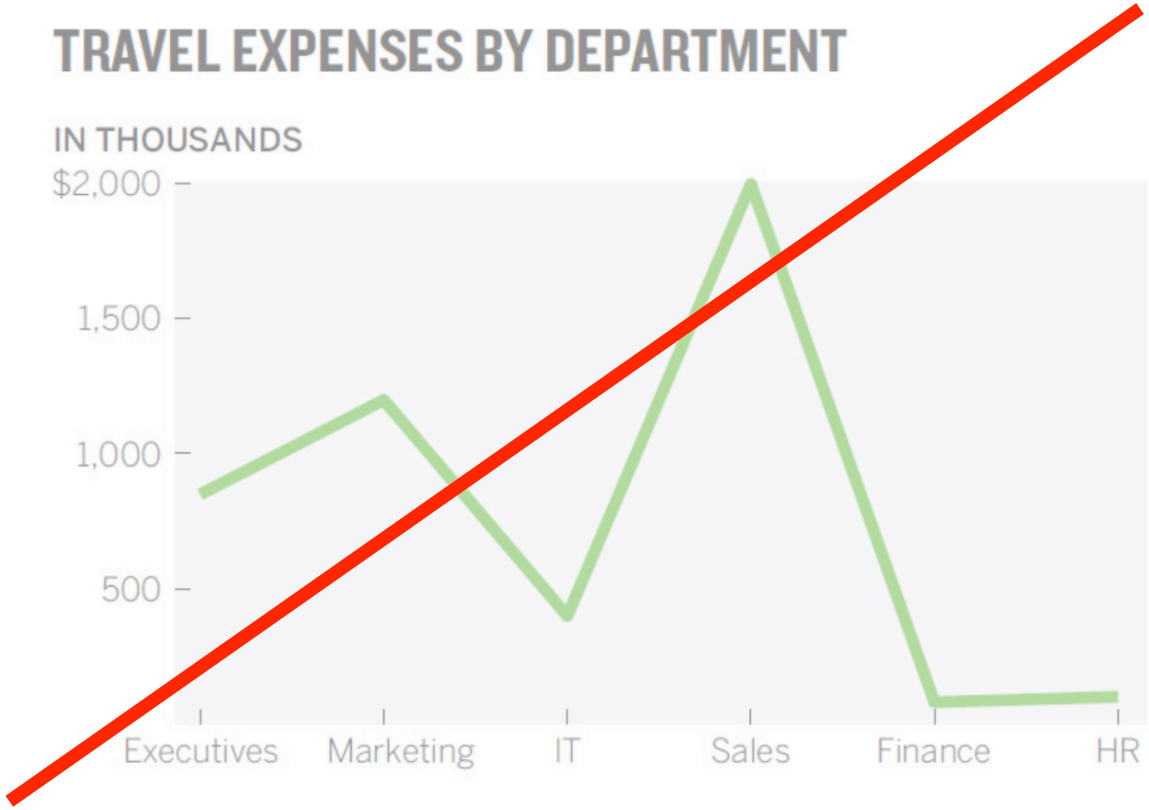
Color: temperature (O) *repeat

OTHERWISE…



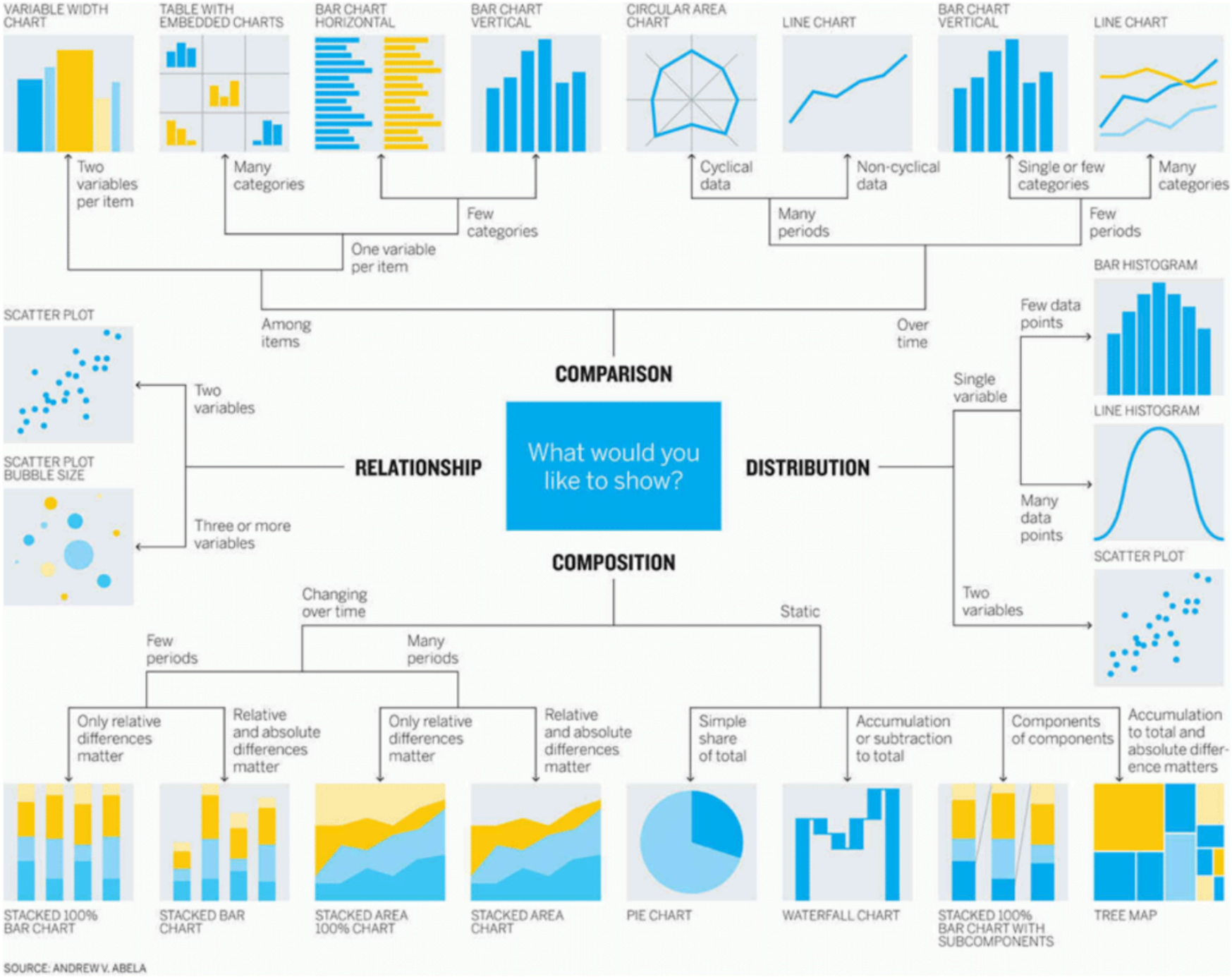
Bernato, *Good Charts* (2016).

DON'T DO THIS


TRAVEL EXPENSES BY DEPARTMENT
IN THOUSANDS
$2,000
1,500
1,000
500
Executives
Marketing
IT
Sales
Finance
HR

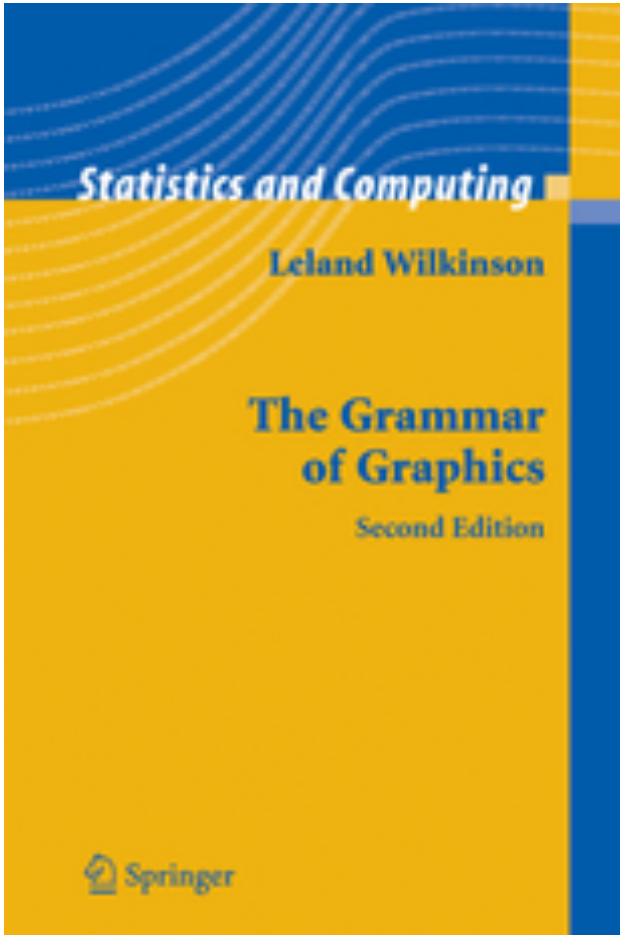
WHICH VIZ TYPE FOR WHICH TASK?

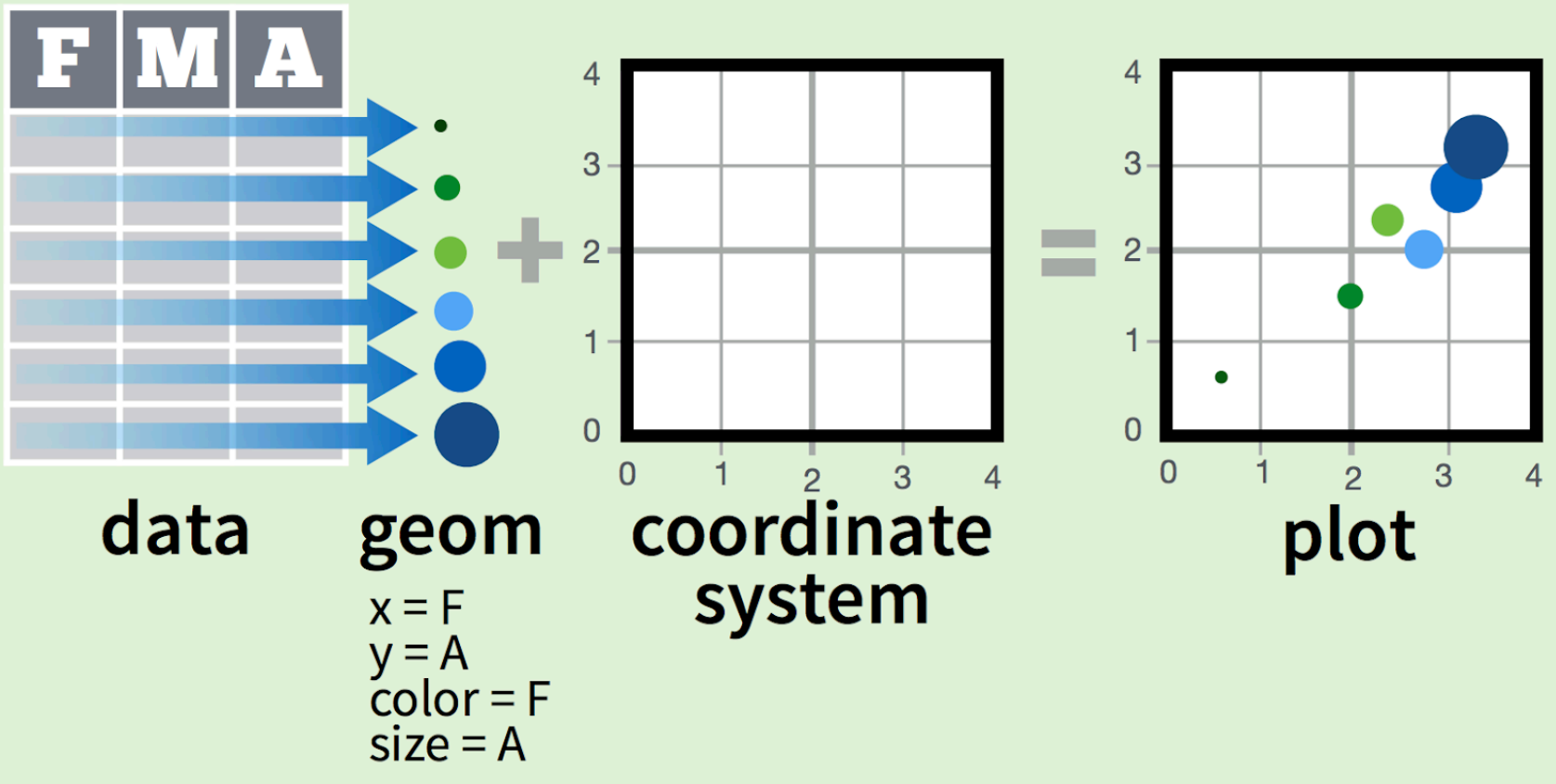
- Depends on your data types;
- But mostly depends on what relationship you want to show.
- No single visualization can express everything perfectly; instead you must prioritize and choose where to focus your reader's attention.



SOURCE: ANDREW V. ABELA

GRAMMAR OF GRAPHICS




F
M
A
0
1
2
3
4
data
geom
x = F
y = A
color = F
size = A
coordinate system
plot

VISUALIZATION GRAMMAR

Data	*Input data to visualize*
Transform	*Grouping, stats, projection*
Mark	*Data-representative geometry*
Encoding	*Mapping between data and mark*
Scale	*Map data values to visual values*
Guide	*Axes & legends to describe scales*

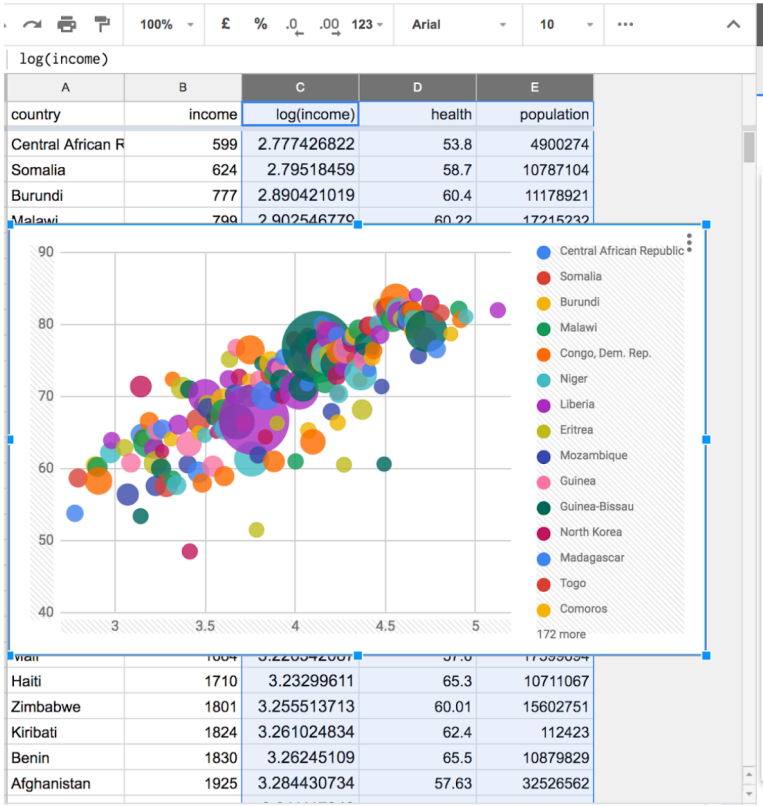

100%
£
%
.0
.00
123
Arial
10
log(income)
A
B
C
D
E
country
income
log(income)
health
population
Central African R
599
2.777426822
53.8
4900274
Somalia
624
2.79518459
58.7
10787104
Burundi
777
2.890421019
60.4
11178921
Haiti
1710
3.23299611
65.3
10711067
Zimbabwe
1801
3.255513713
60.01
15602751
Kiribati
1824
3.261024834
62.4
112423
Benin
1830
3.26245109
65.5
10879829
Afghanistan
1925
3.284430734
57.63
32526562
90
80
70
60
50
40
3
3.5
4
4.5
5
Central African Republic
Somalia
Burundi
Malawi
Congo, Dem. Rep.
Niger
Liberia
Eritrea
Mozambique
Guinea
Guinea-Bissau
North Korea
Madagascar
Togo
Comoros
172 more


Chart editor
DATA
CUSTOMIZE
Chart type
Bubble chart
Line
Area
Column
Bar
Pie

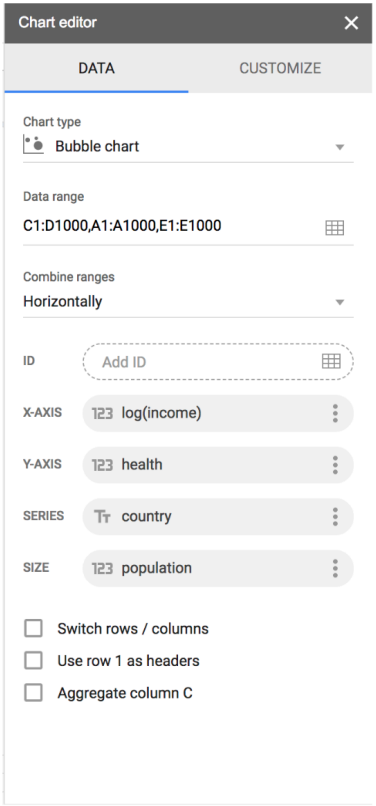


Chart editor
DATA
CUSTOMIZE
Chart type
Bubble chart
Data range
C1:D1000,A1:A1000,E1:E1000
Combine ranges
Horizontally
ID
Add ID
X-AXIS
log(income)
Y-AXIS
health
SERIES
country
SIZE
population
Switch rows / columns
Use row 1 as headers
Aggregate column C

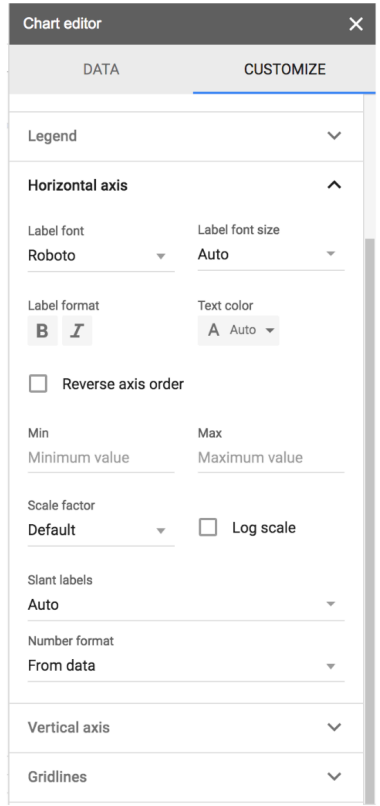
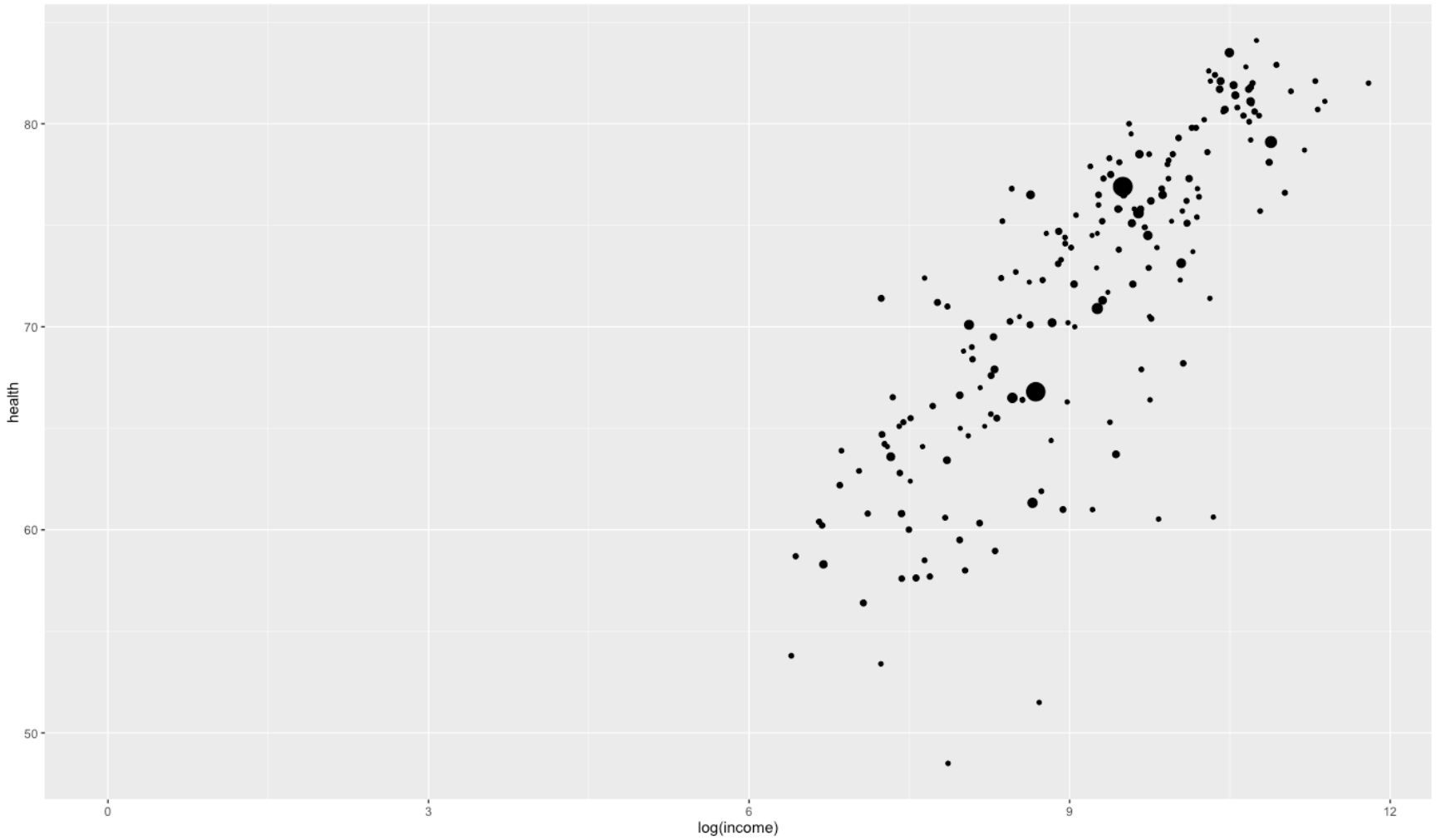


Chart editor
DATA
CUSTOMIZE
Legend
Horizontal axis
Label font
Roboto
Label font size
Auto
Label format
Text color
Auto
Reverse axis order
Min
Minimum value
Max
Maximum value
Scale factor
Default
Log scale
Slant labels
Auto
Number format
From data
Vertical axis
Gridlines




```
ggplot(d) +
  geom_point(aes(x=log(income), y=health, size=population))
```

FUN THINGS TO READ

- *Visual Explanations*, e.g. Chp. 2 Challenger shuttle, by E. Tufte
- *A Tour through the Visualization Zoo* by J. Heer
- *39 Studies About Human Perception in 30 Minutes* by K. Elliot
- *Bad Data Guide* by Quartz data team

LAB INSTRUCTIONS

Open assignment: https://classroom.github.com/a/1AJHn-vY

Click “Clone or Download” to get repository URL

In terminal, type git clone [URL] to get a local copy

Start RStudio, open `3-ggplot-exercises.Rmd`