

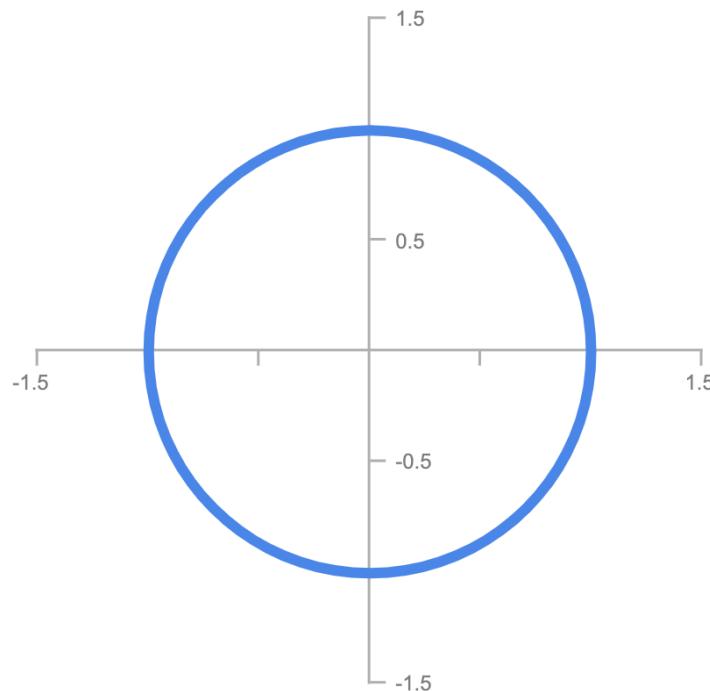
VISUALIZATION IN R WITH GG PLOT 2

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?



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

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

$$\mu_X = 9.0, \sigma_X = 11$$

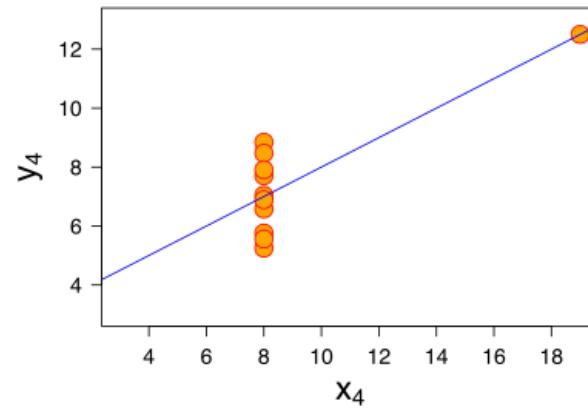
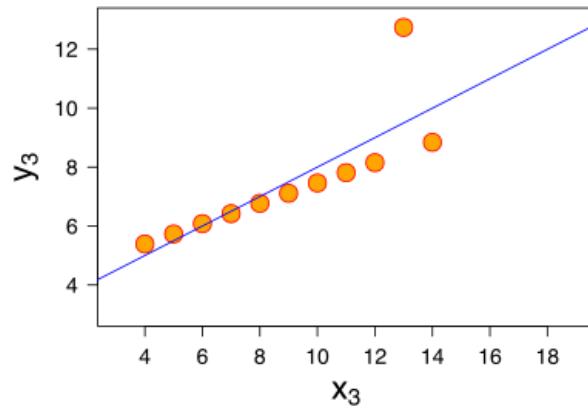
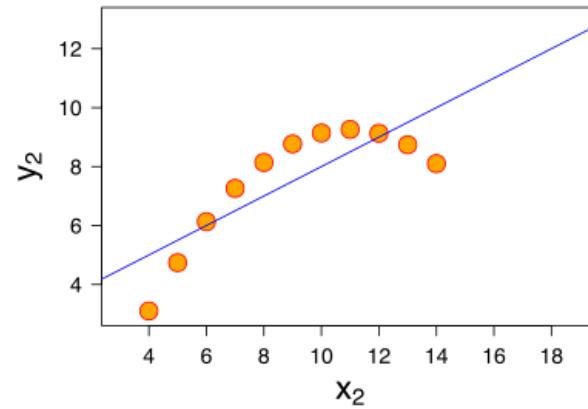
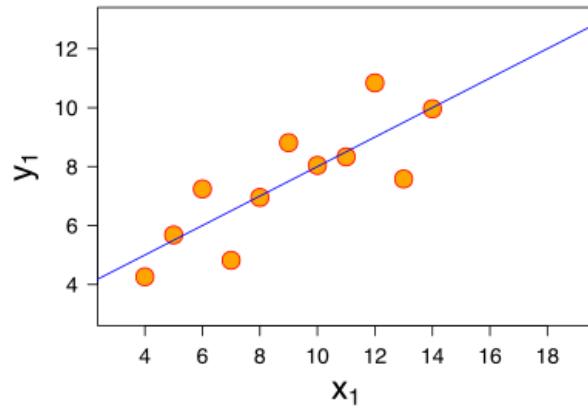
$$\mu_Y = 7.5, \sigma_Y = 4.125$$

Linear Regression

$$Y = 3 + 0.5X$$

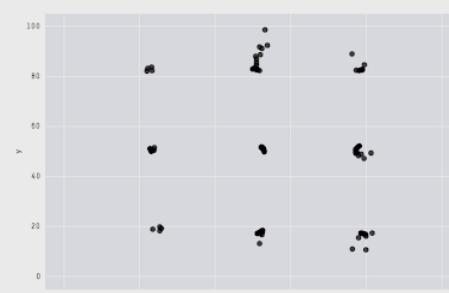
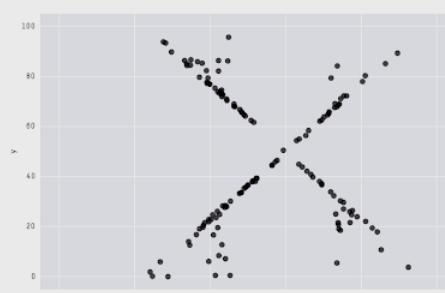
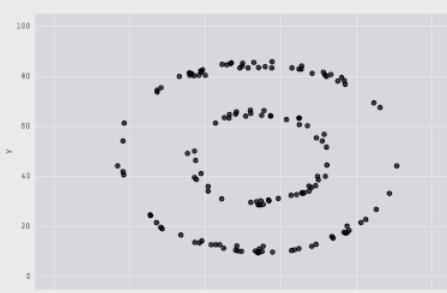
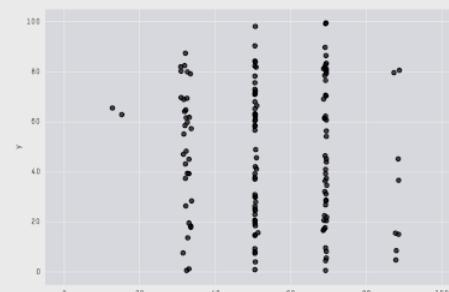
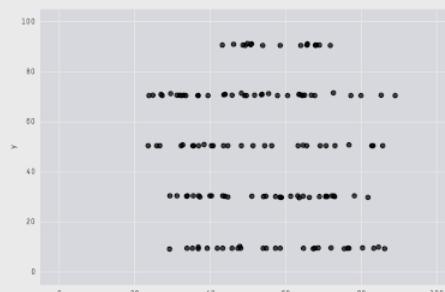
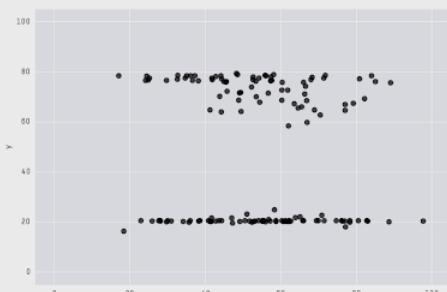
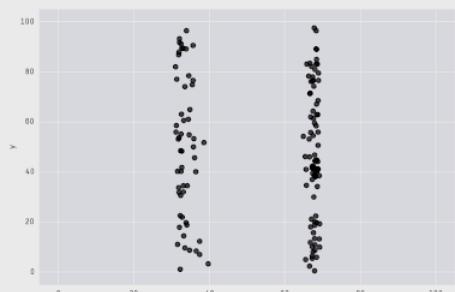
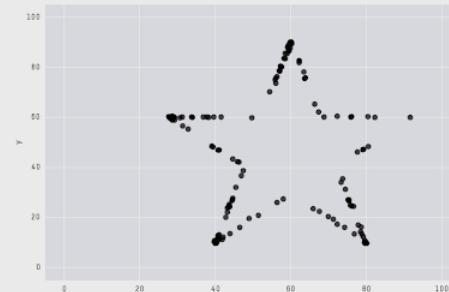
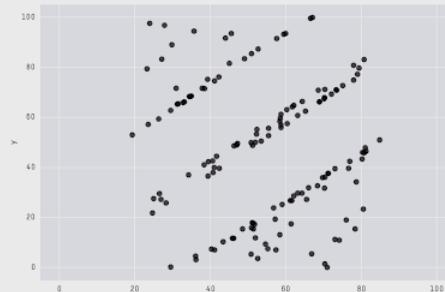
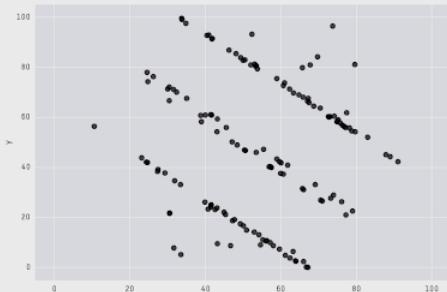
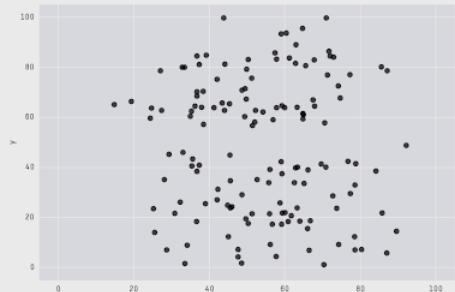
$$R^2 = 0.67$$

ANScombe's Quartet

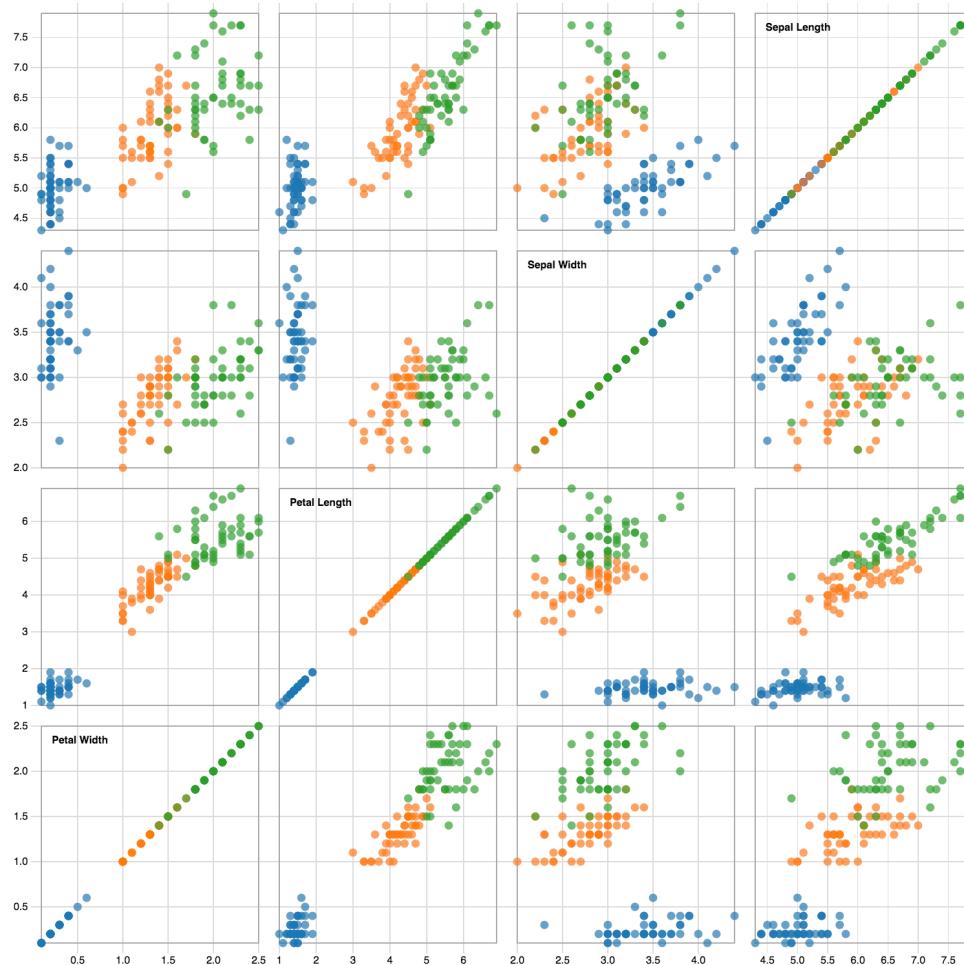




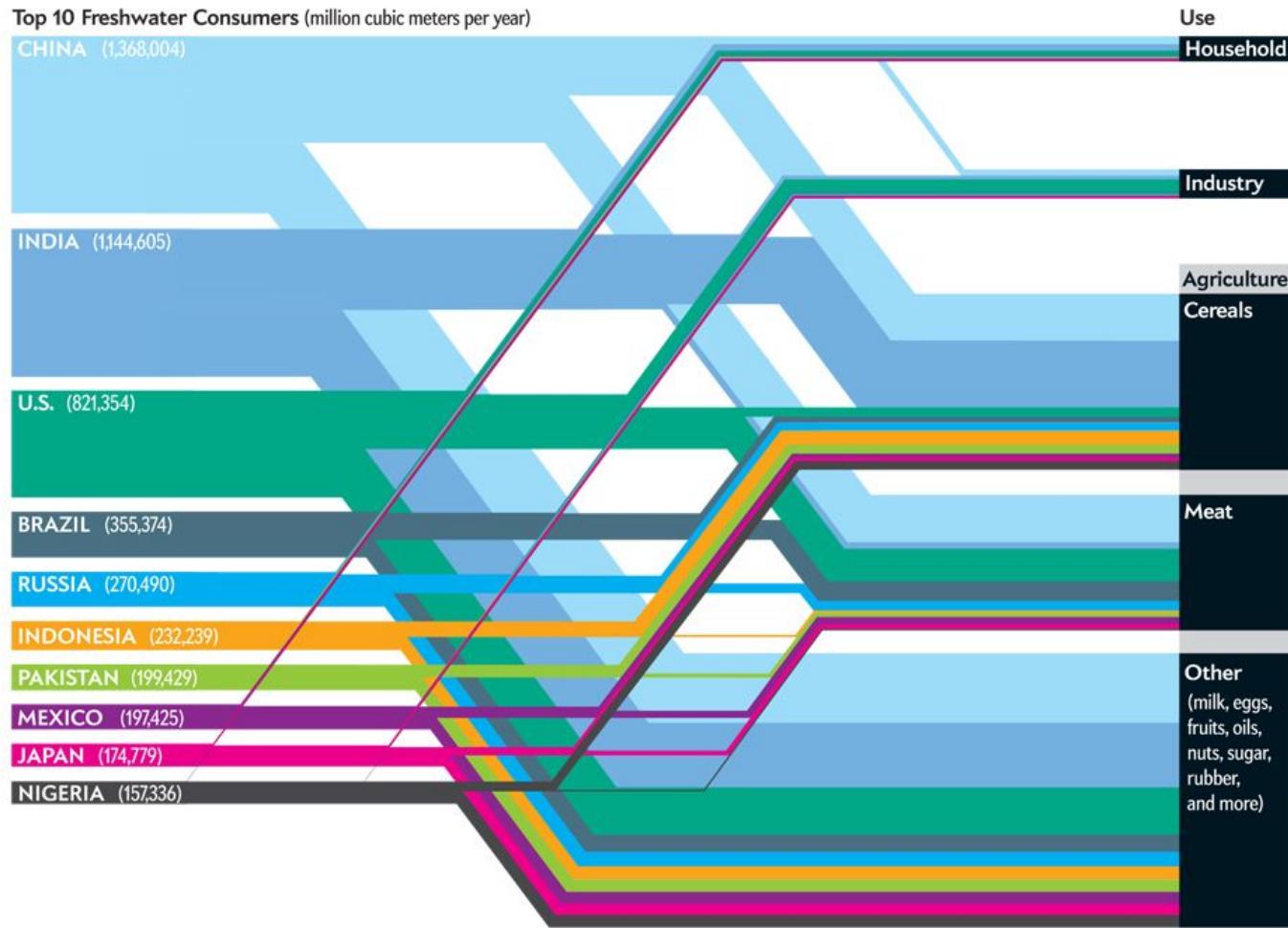
X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06



EXPLORATORY: “SPLOM”



EXPLANATORY: SCI. AM.



GOALS OF VISUALIZATION

Exploratory

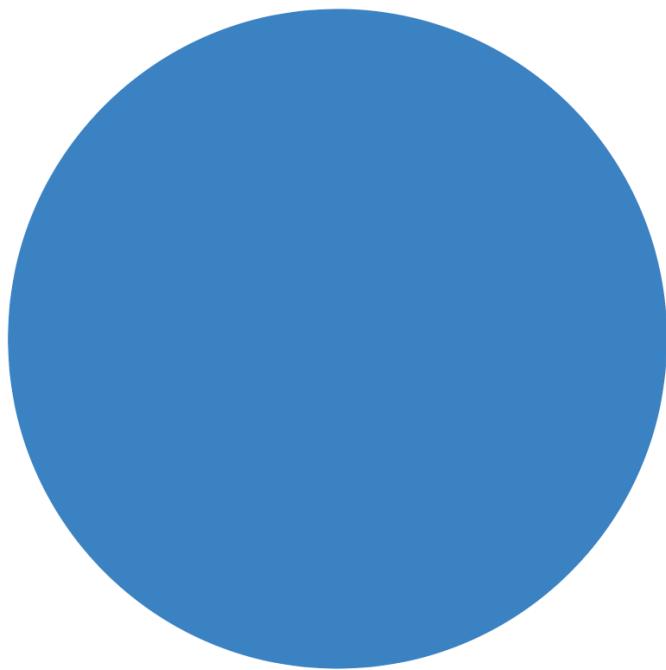
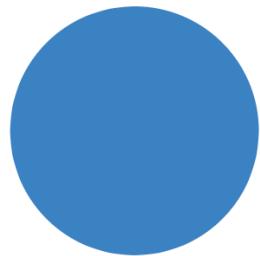
- to uncover a relationship in the data
- to analyze data

Explanatory

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

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

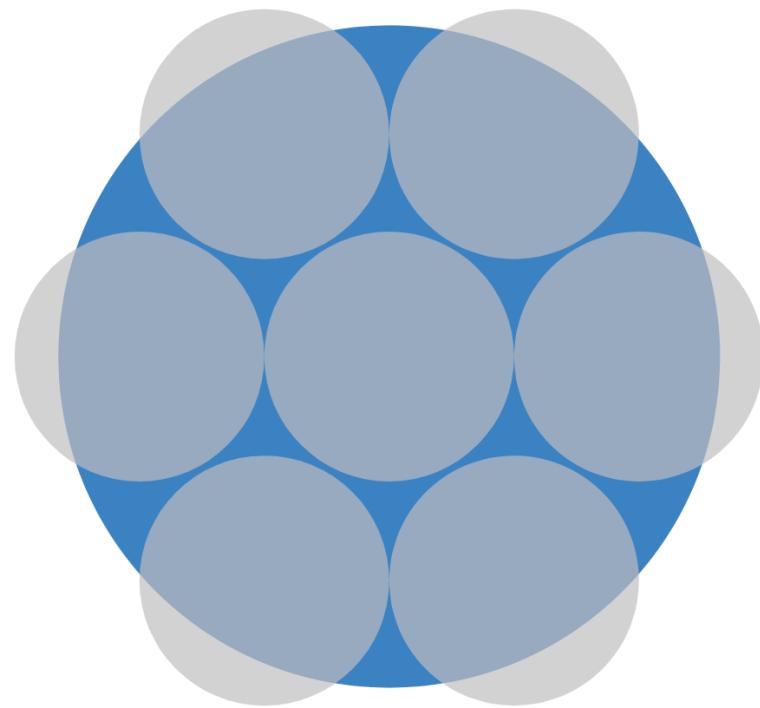
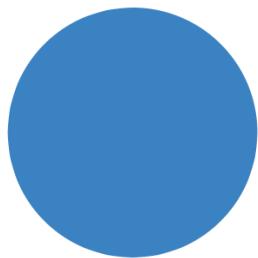
COMPARE AREA OF CIRCLES



COMPARE LENGTH OF BARS

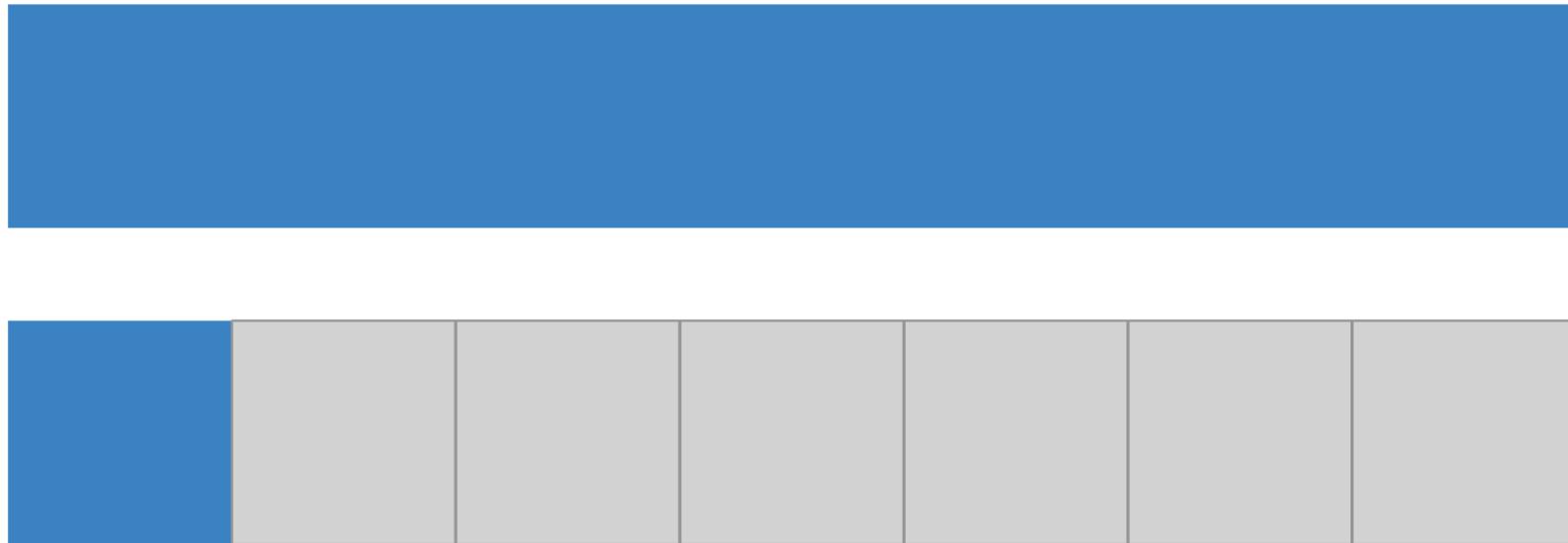


COMPARE AREA OF CIRCLES

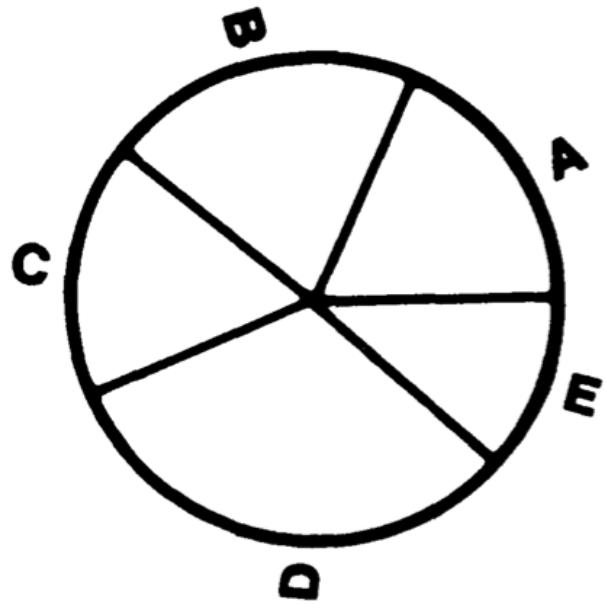


Via Jeff Heer

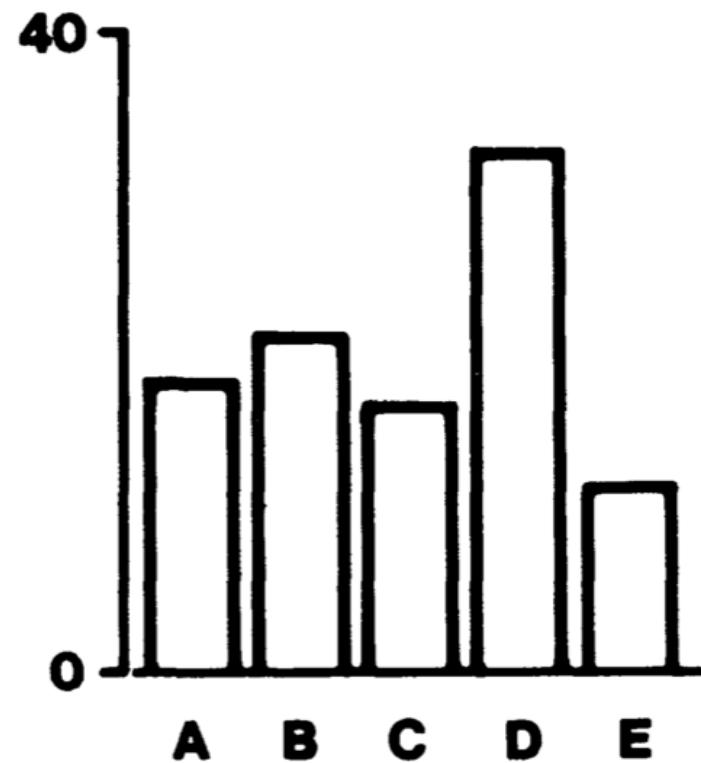
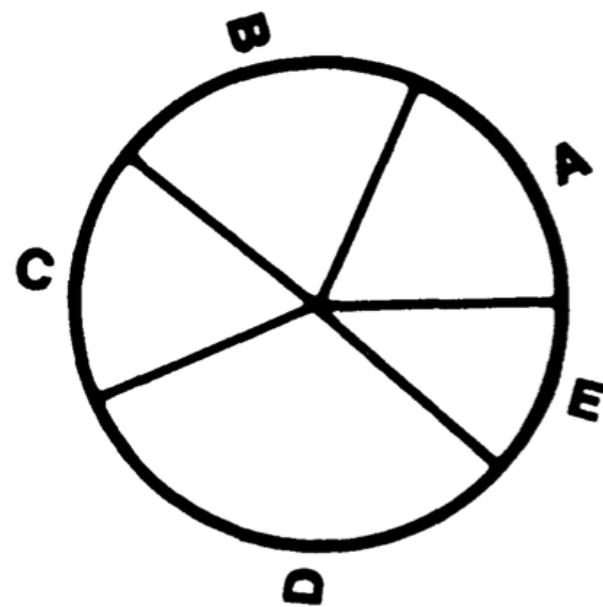
COMPARE LENGTH OF BARS



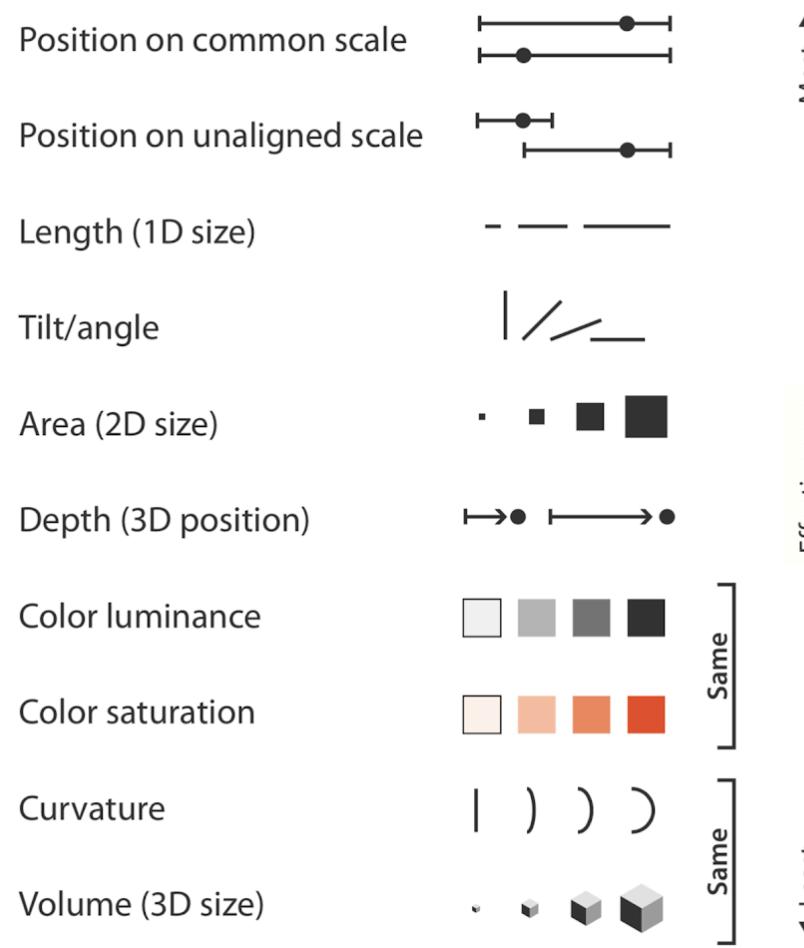
WHICH IS LARGER, A OR C?



WHICH IS LARGER, A OR C?



PERCEPTION RANKING

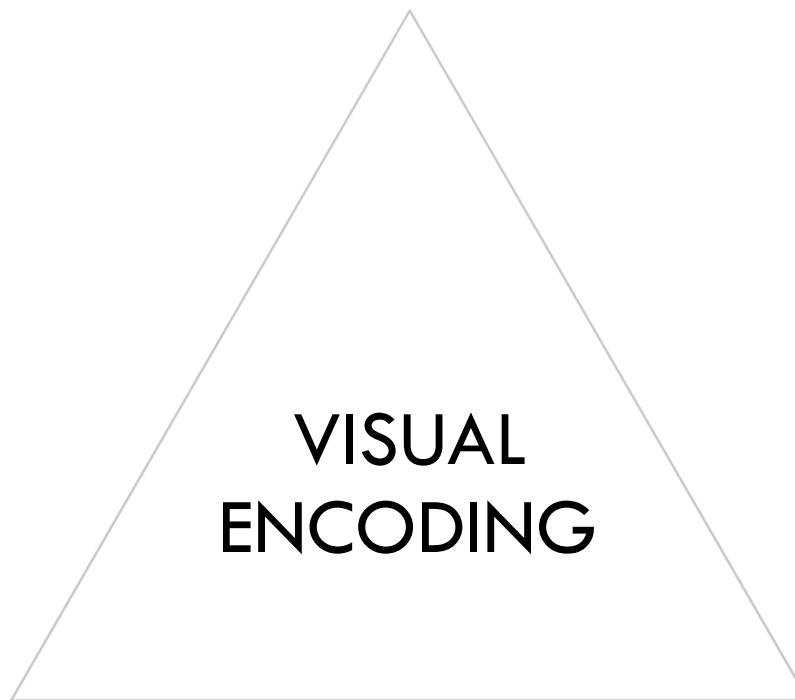
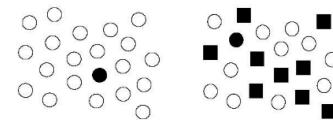


WHAT IS DATA VIZ?

- 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.*

Perceptual Properties



Marks &
Channels



Data Types



MARKS

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

Point



Line



Area



CHANNELS

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

Position



Color



Shape



Tilt



Size

→ Length



→ Area



→ Volume



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

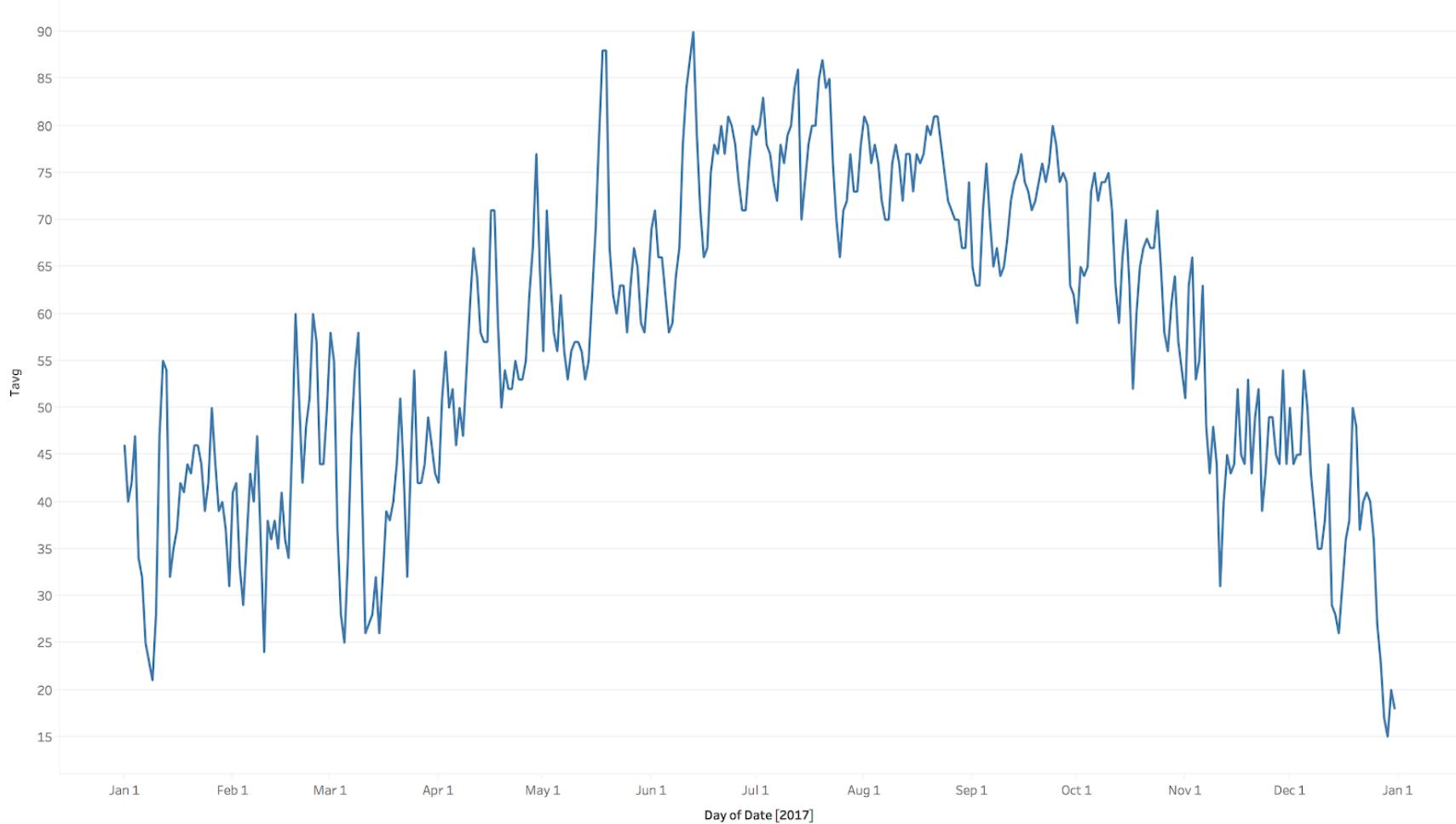
Data Model: 21, 28, 47, 55, ... (integers)

Conceptual Model: Temperature ($^{\circ}\text{F}$)

Data Type:

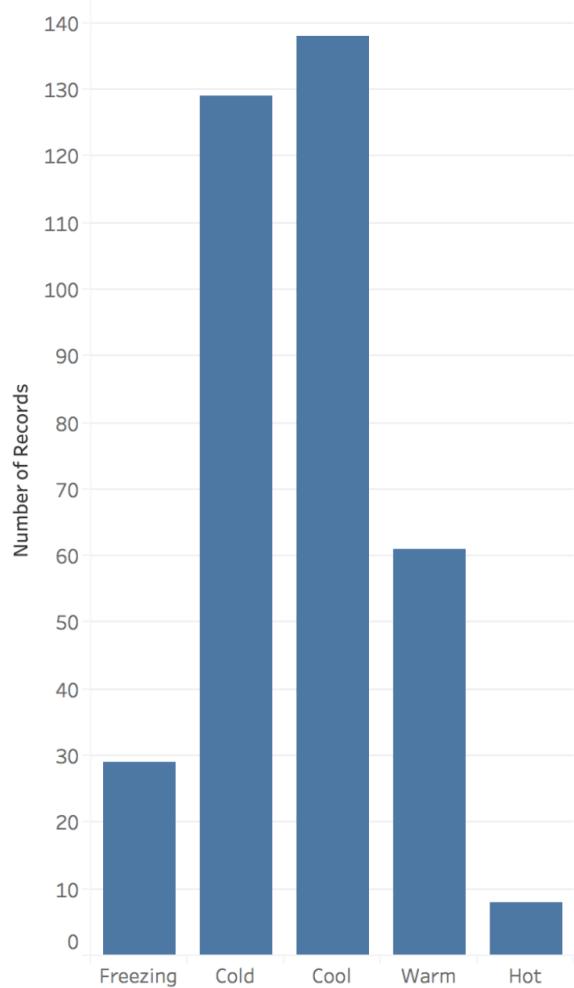
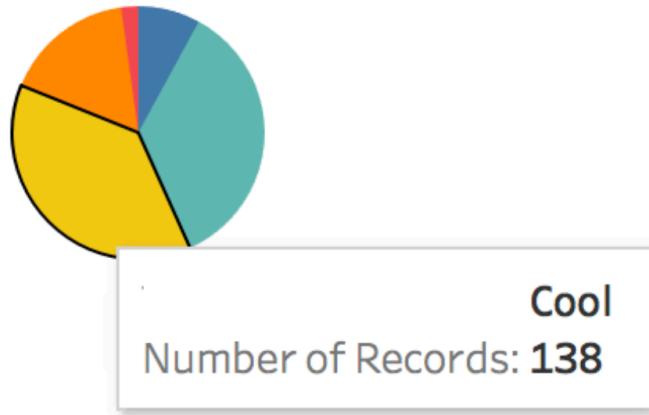
- Freezing vs. Not-Freezing (C)
- Hot, Warm, Cold, Freezing (O)
- Temperature Value (Q)

E.G. QUANT X QUANT



E.G. ORDINAL X QUANT

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



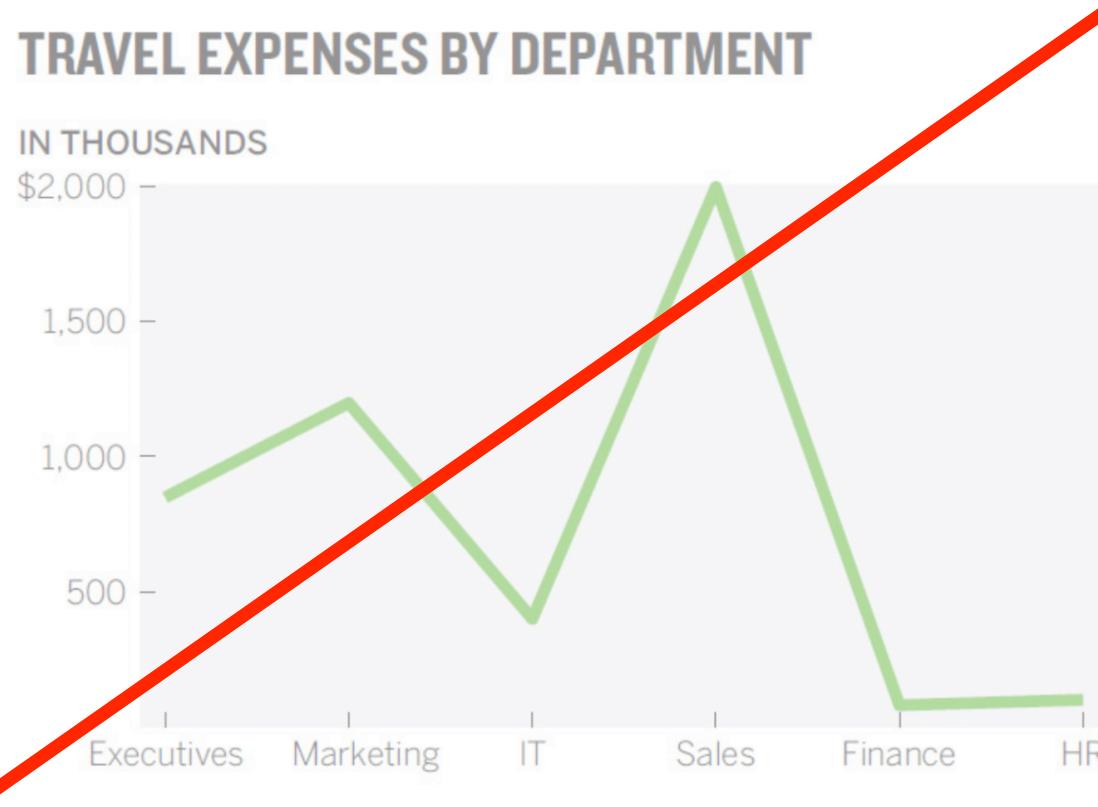
OTHERWISE...

TRAVEL EXPENSES BY DEPARTMENT

IN THOUSANDS

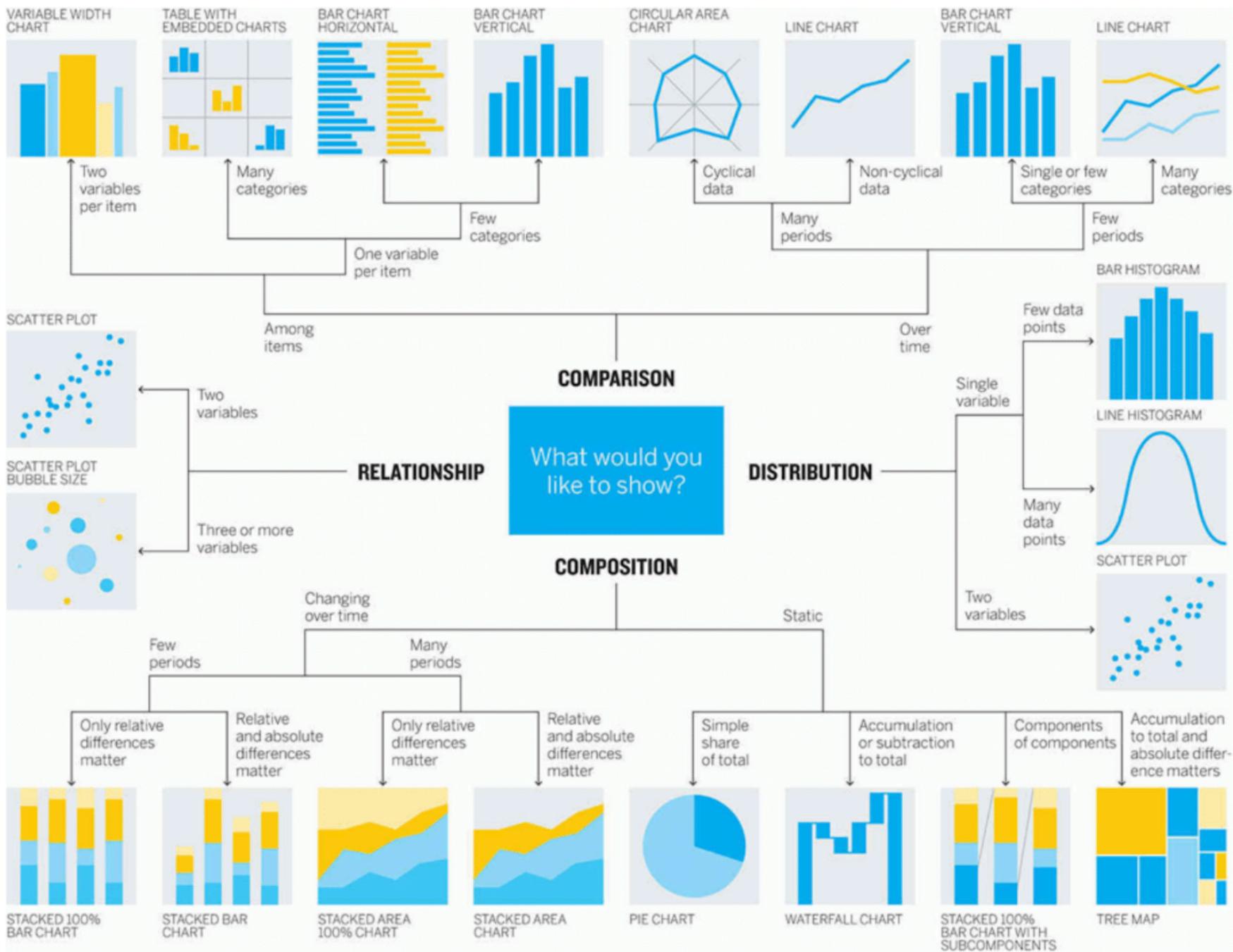


DON'T DO THIS

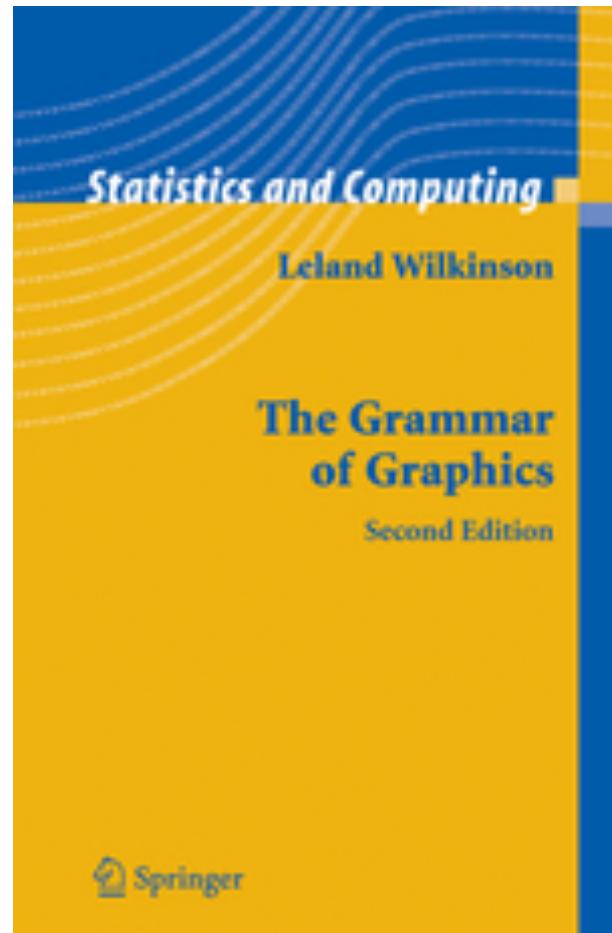


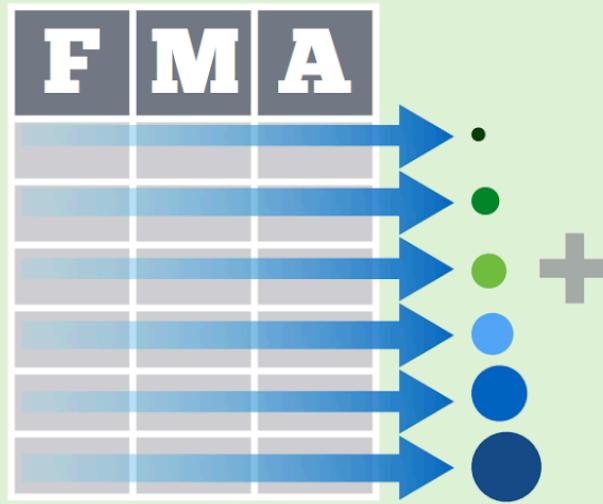
WHICH VIZ TYPE FOR WHICH TASK?

- 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.



GRAMMAR OF GRAPHICS

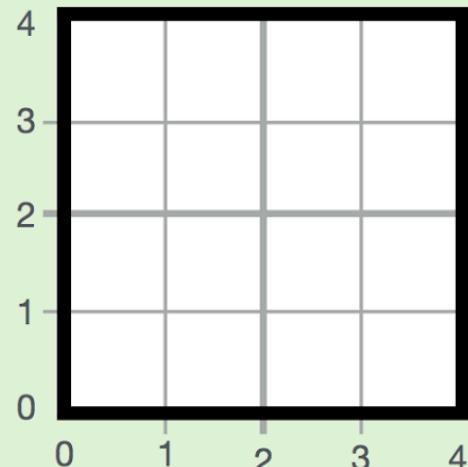




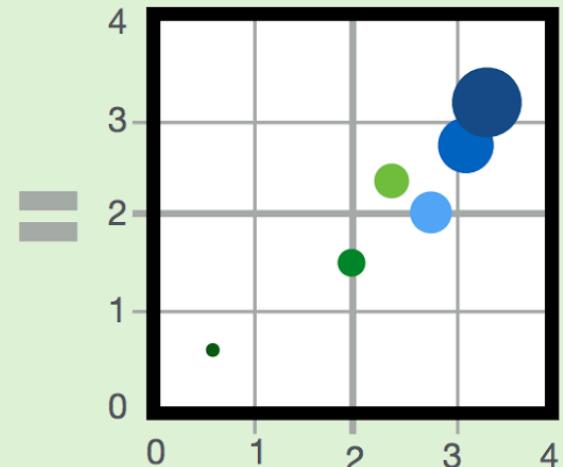
data

geom

$x = F$
 $y = A$
color = F
size = A



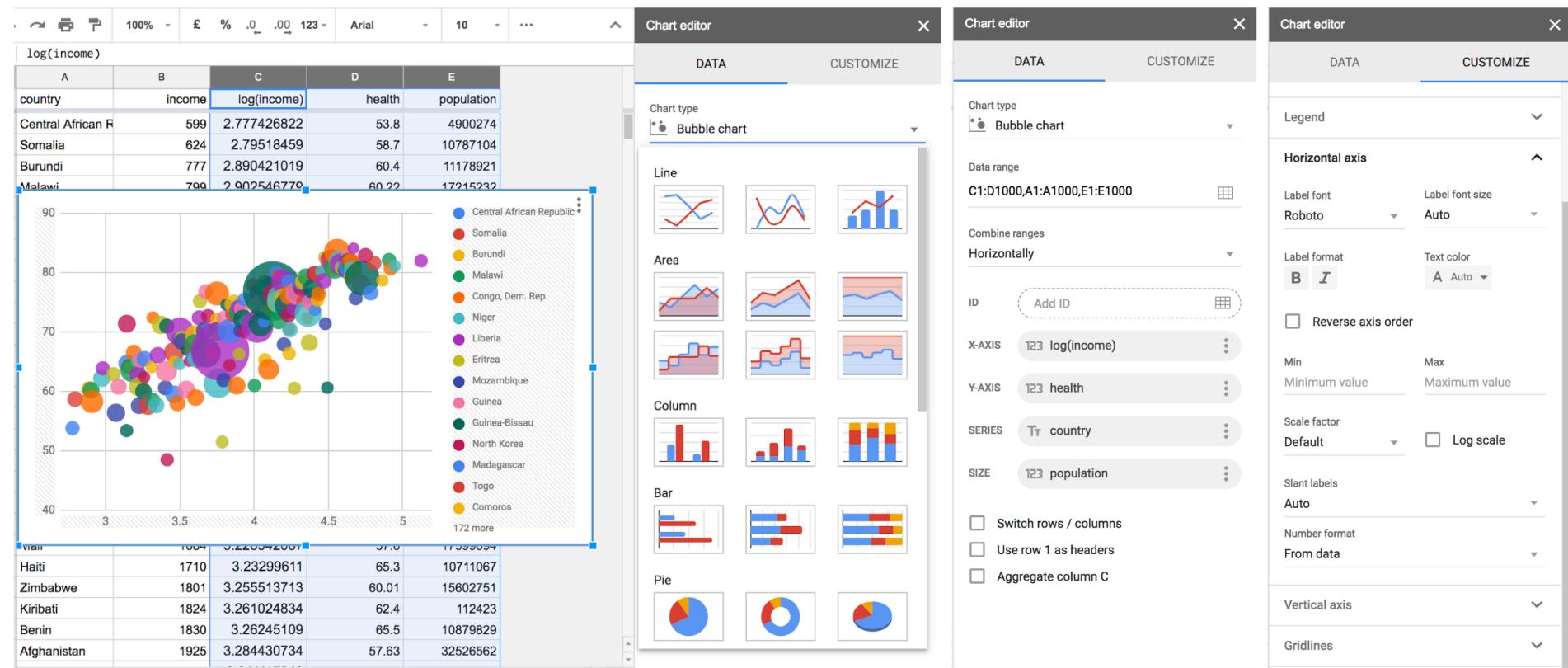
**coordinate
system**

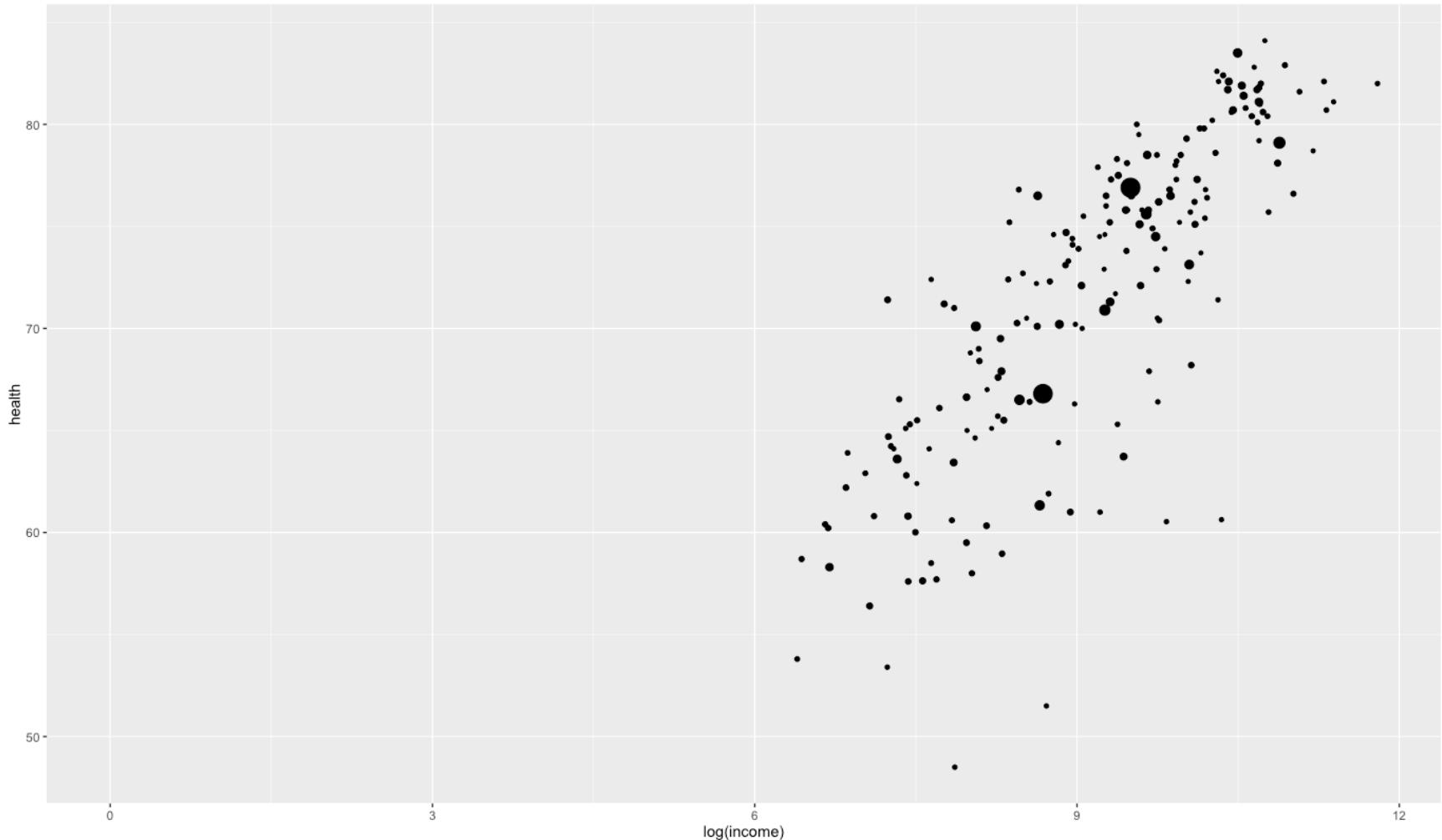


plot

VISUALIZATION GRAMMAR

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





```
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`