Winter 2018-2019 Data Science Bootcamp

Exploratory Data Analysis and Visualization with ggplot2

Joyce Robbins

Lecturer, Columbia University

Dept of Statistics

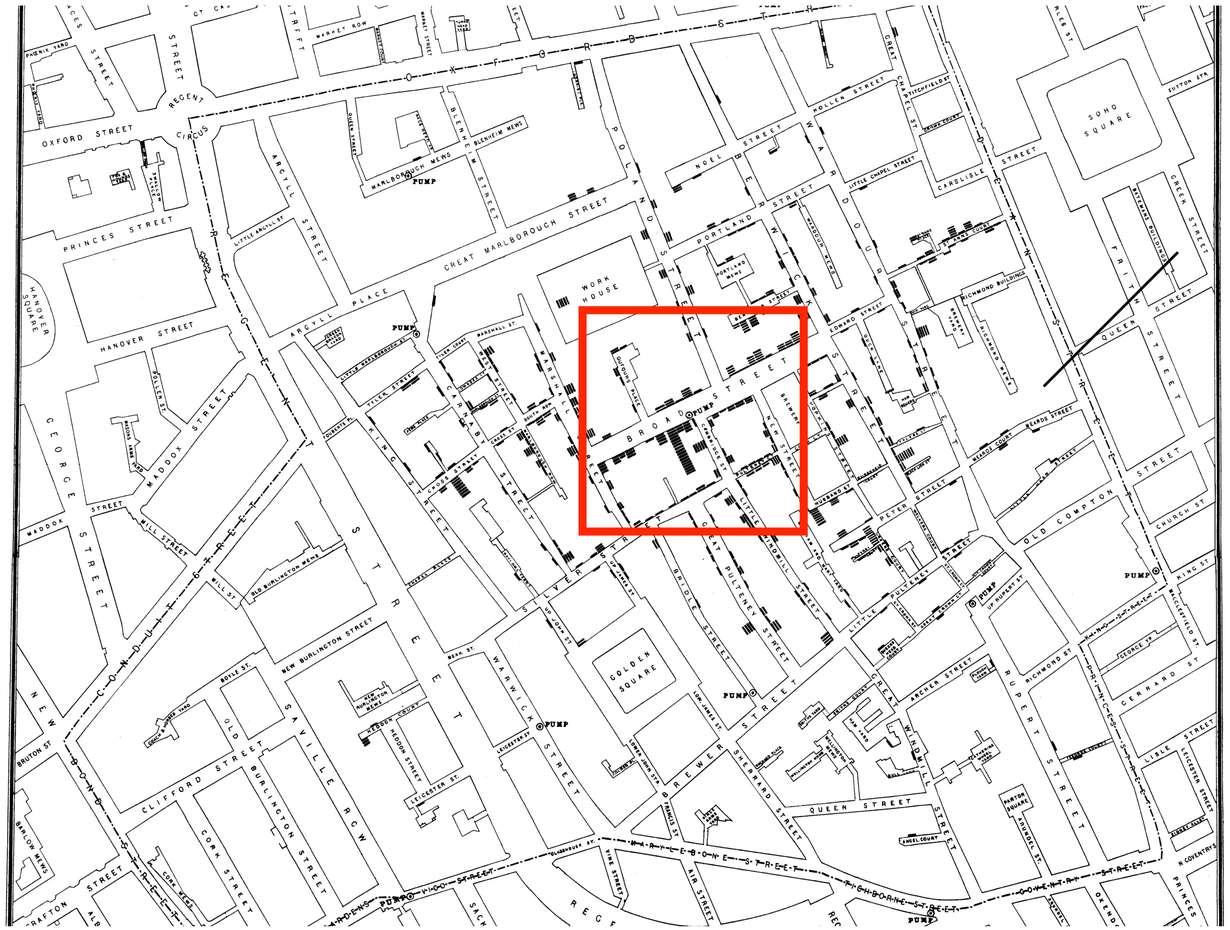
Data Science Institute

EXPLORATORY DATA ANALYSIS

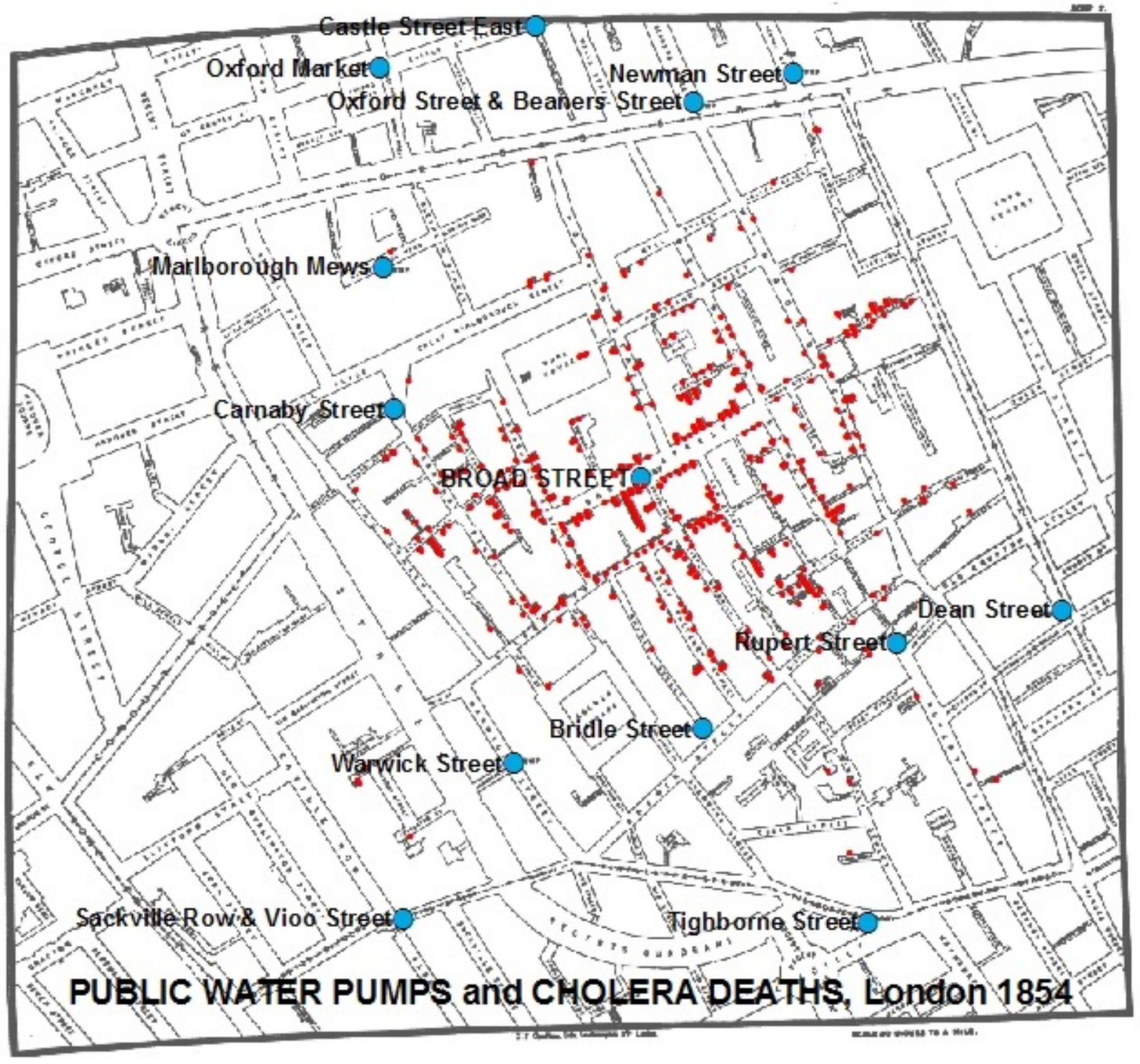
Exploratory Data Analysis

- detecting patterns
- finding outliers
- making comparisons
- identifying clusters

John Snow, Cholera Map 1854



Broad St. pump


Castle Street East
Oxford Market
Newman Street
Oxford Street & Beaners Street
Marlborough Mews
Carnaby Street
BROAD STREET
Dean Street
Rupert Street
Bridle Street
Warwick Street
Sackville Row & Vioo Street
Tighborne Street
PUBLIC WATER PUMPS and CHOLERA DEATHS, London 1854

Anscombe's Quartet

x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.10	5.39	12.50
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

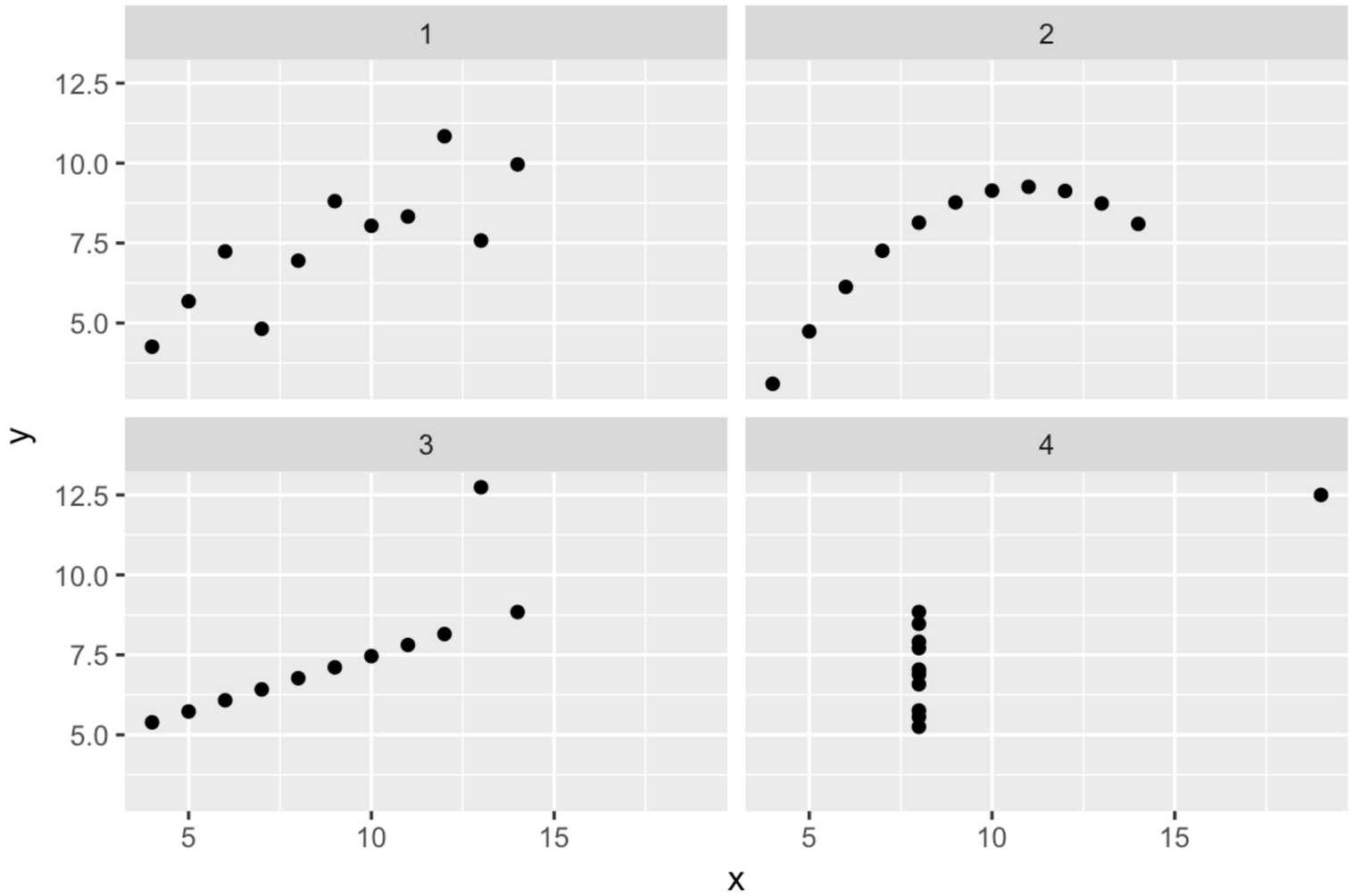
Numeric summary

Each of the four data sets yields the same standard output from a typical regression program, namely

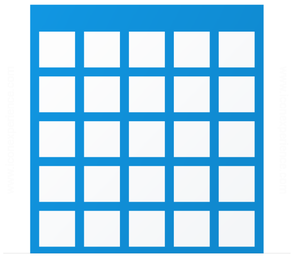
Number of observations $(n) = 11$
Mean of the x's $(\bar{x}) = 9.0$
Mean of the y's $(\bar{y}) = 7.5$
Regression coefficient (b_1) of y on $x = 0.5$
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x} = 110.0$
Regression sum of squares $= 27.50$ (1 d.f.)
Residual sum of squares of $y = 13.75$ (9 d.f.)
Estimated standard error of $b_1 = 0.118$
Multiple $R^2 = 0.667$

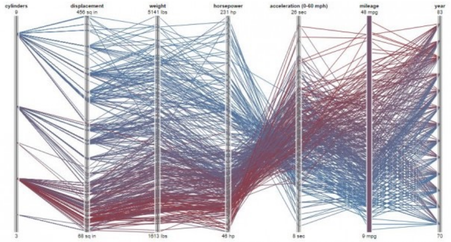
F.J. Anscombe, "Graphs in Statistical Analysis"

Graphical summary



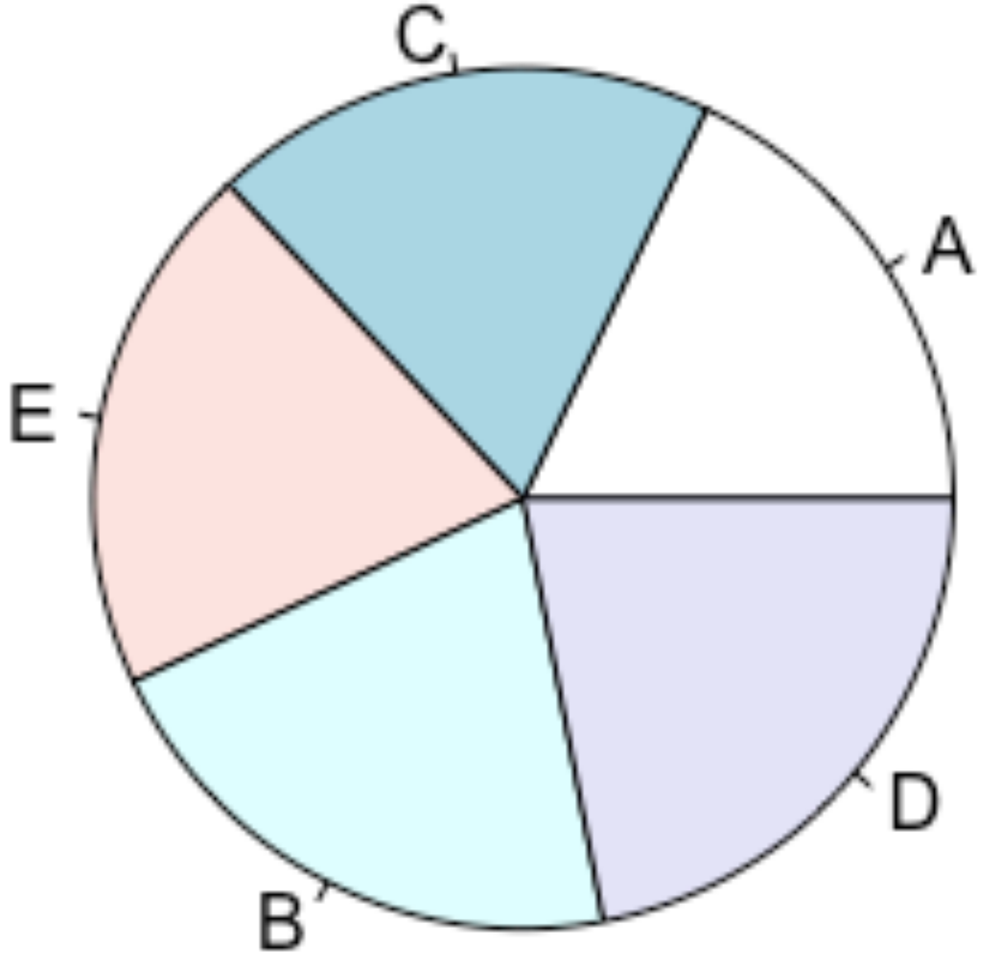
How do we gain insight?



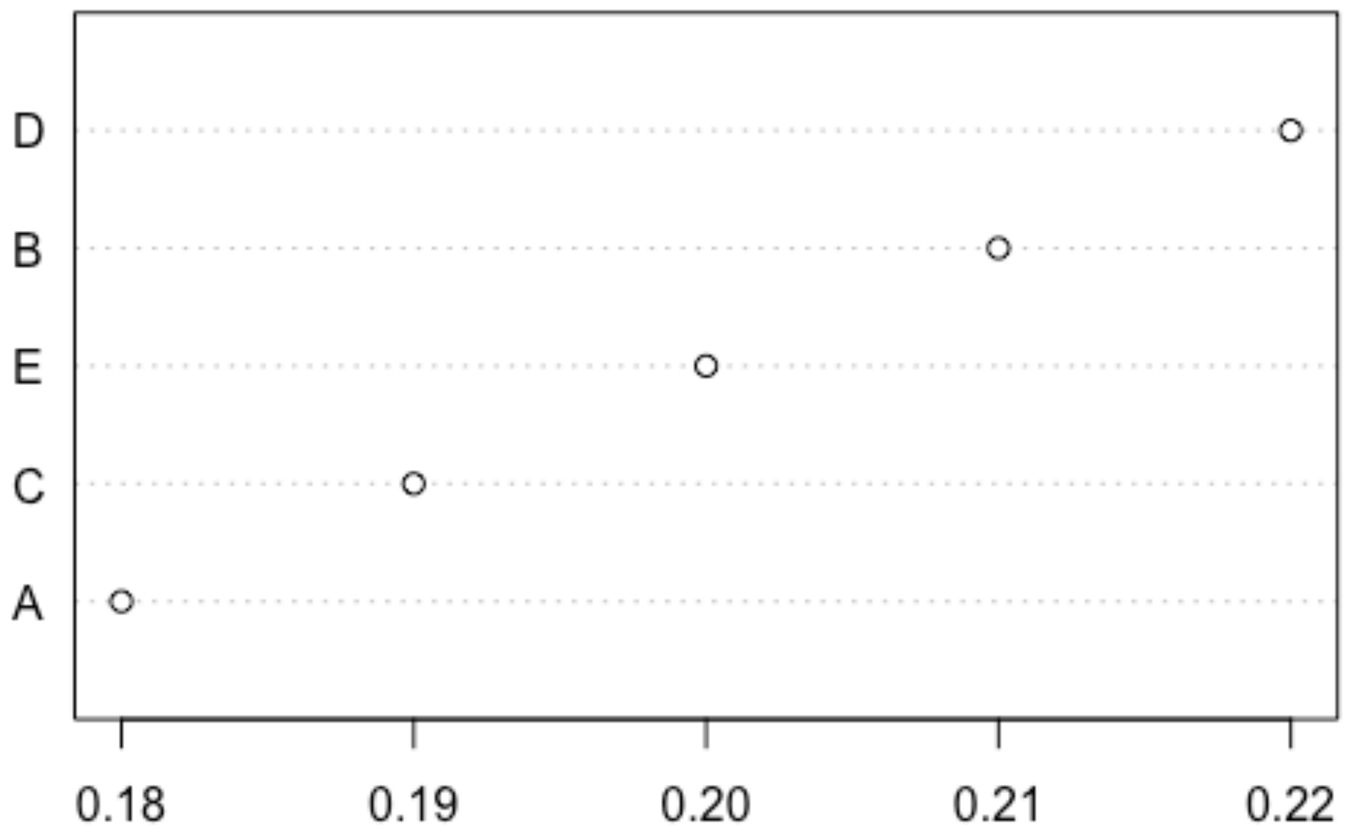


- Deep understanding of the dataset, where it came from, what its limitations are
- Experiment with different graphic forms, based on theory on what forms work well with different data types

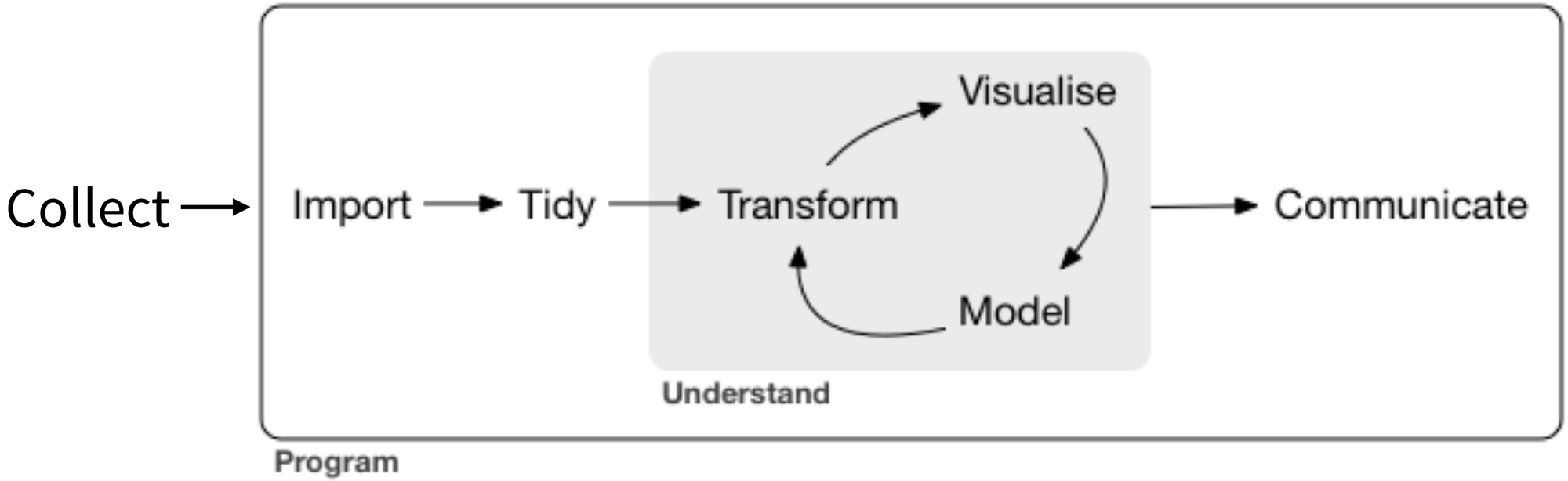
Perception



Perception



Data Science Pipeline



Source: r4ds.had.co.nz/introduction.html

"Visualization is a fundamentally human activity."

VISUALIZATION

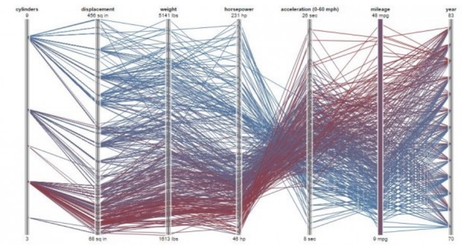
What is data visualization?

- relatively new field (but long history)
- multidisciplinary
- lack of consensus

Interdisciplinary influences

Statistics

Subject Expertise



Graphic Design

Psychology

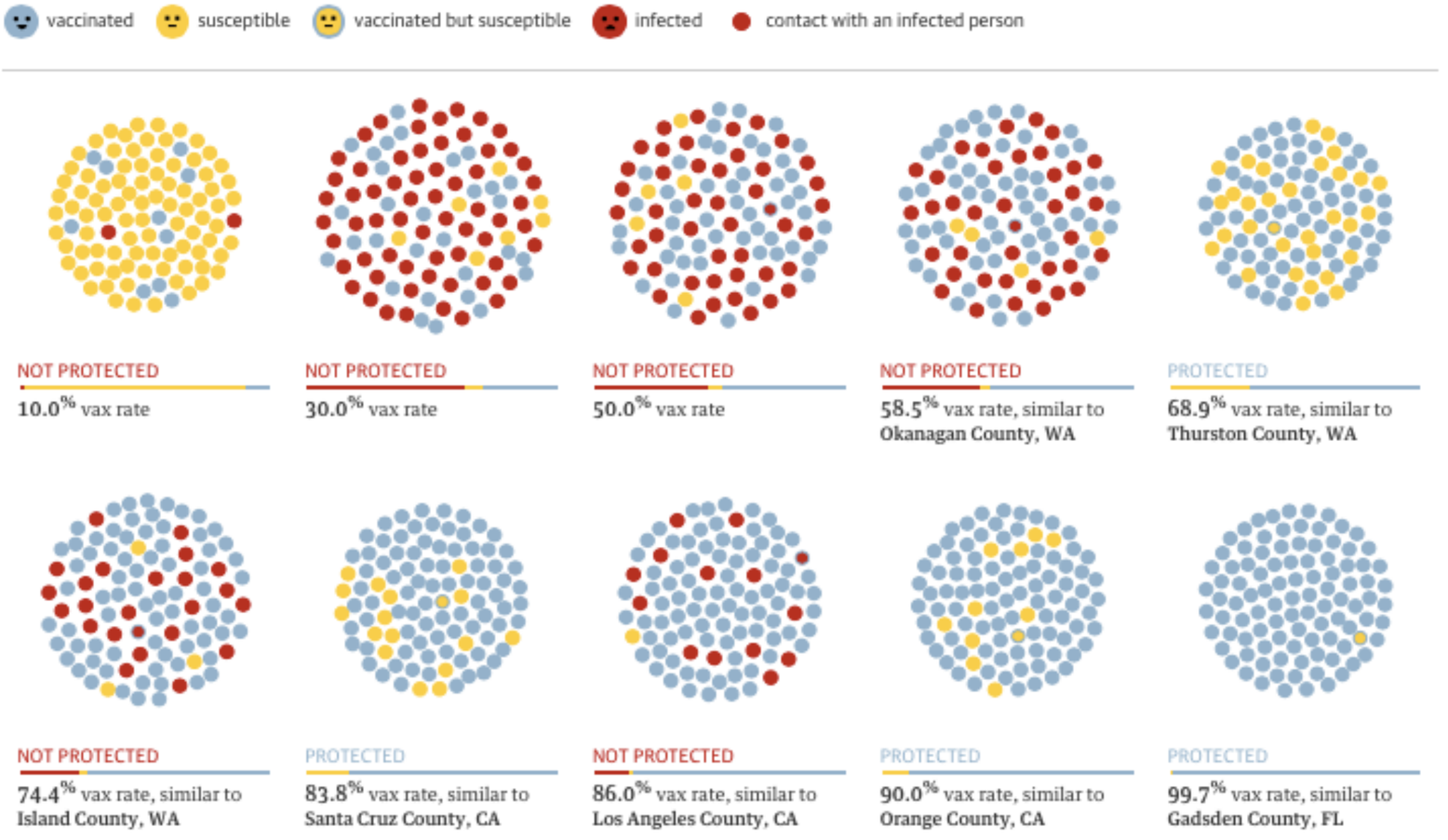
Computer Science

Exploration vs. Visualization (Presentation)

- Sometimes called "exploratory" vs. "explanatory"
- Not mutually exclusive
- Visualizations that offer insight are likely to be shared
- Still, focus is different when exploratory a dataset for the first time vs. presenting to an audience, particularly a less technical one

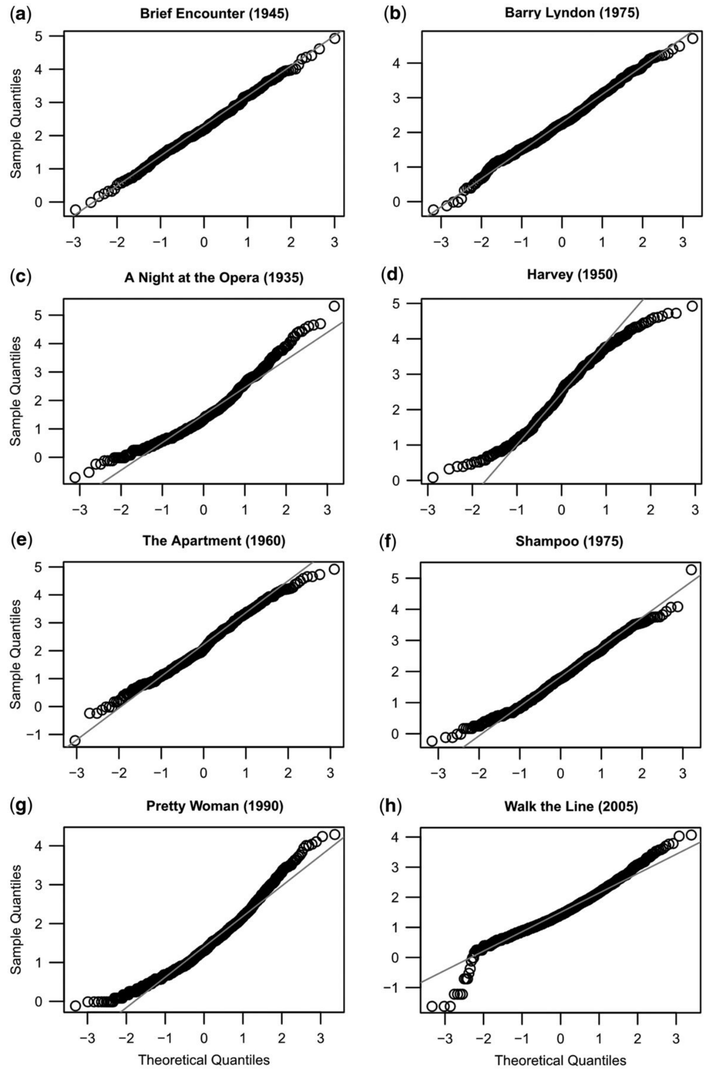
< 55,272

Watch how the measles outbreak spreads when kids get vaccinated – and when they don't



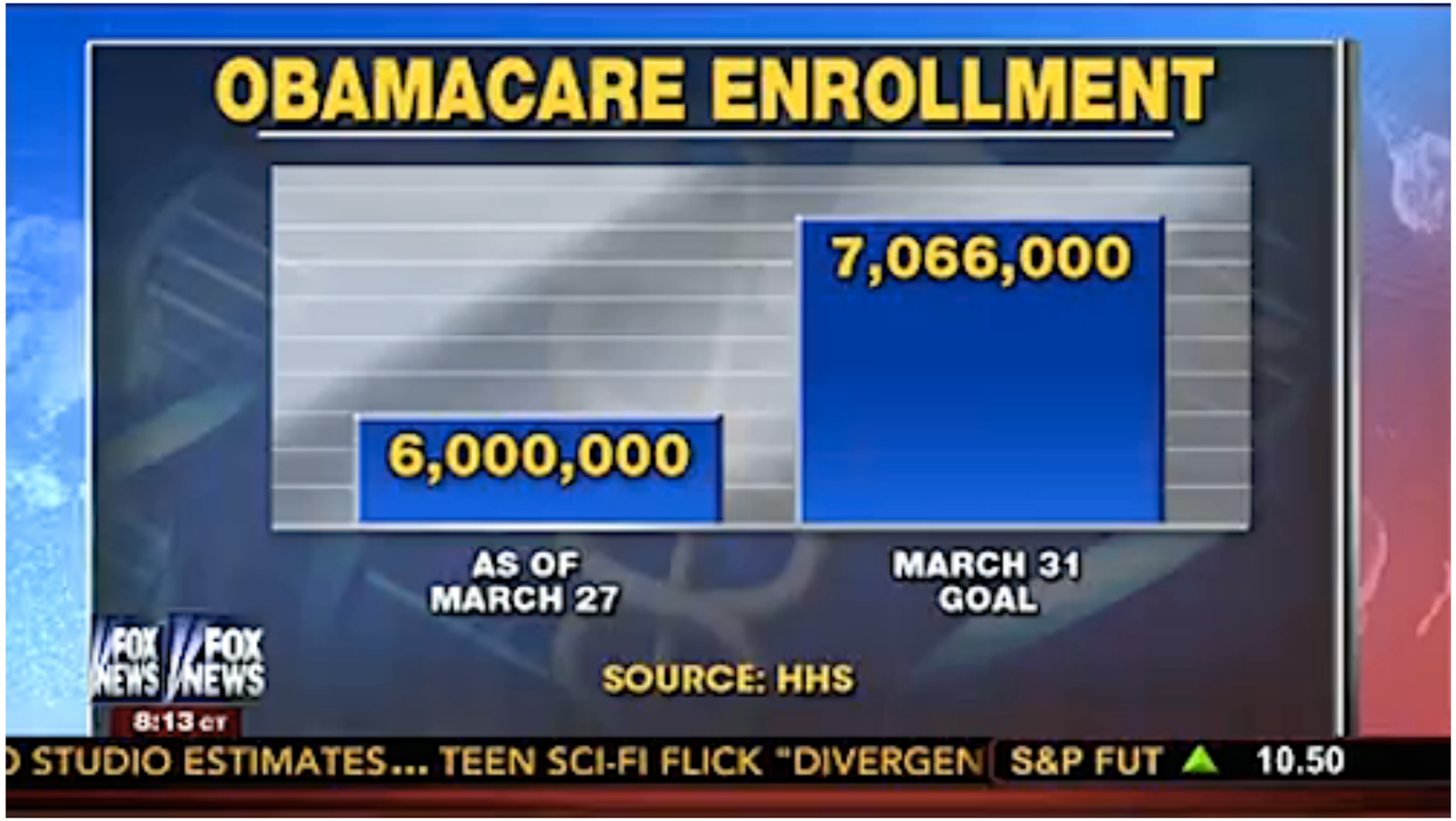
Run simulation again

Audience

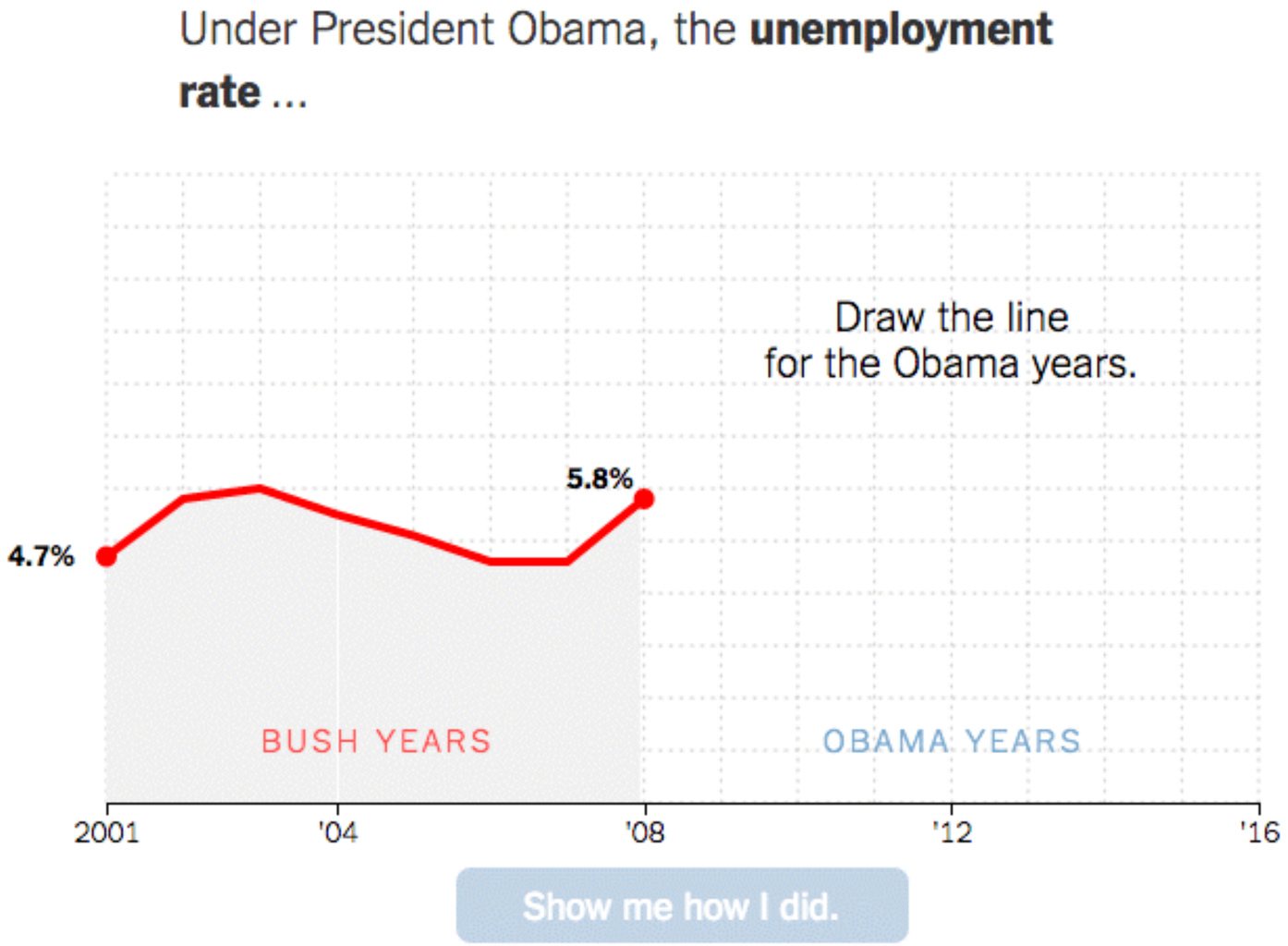


Normal probability plots of log-transformed shot lengths for eight films

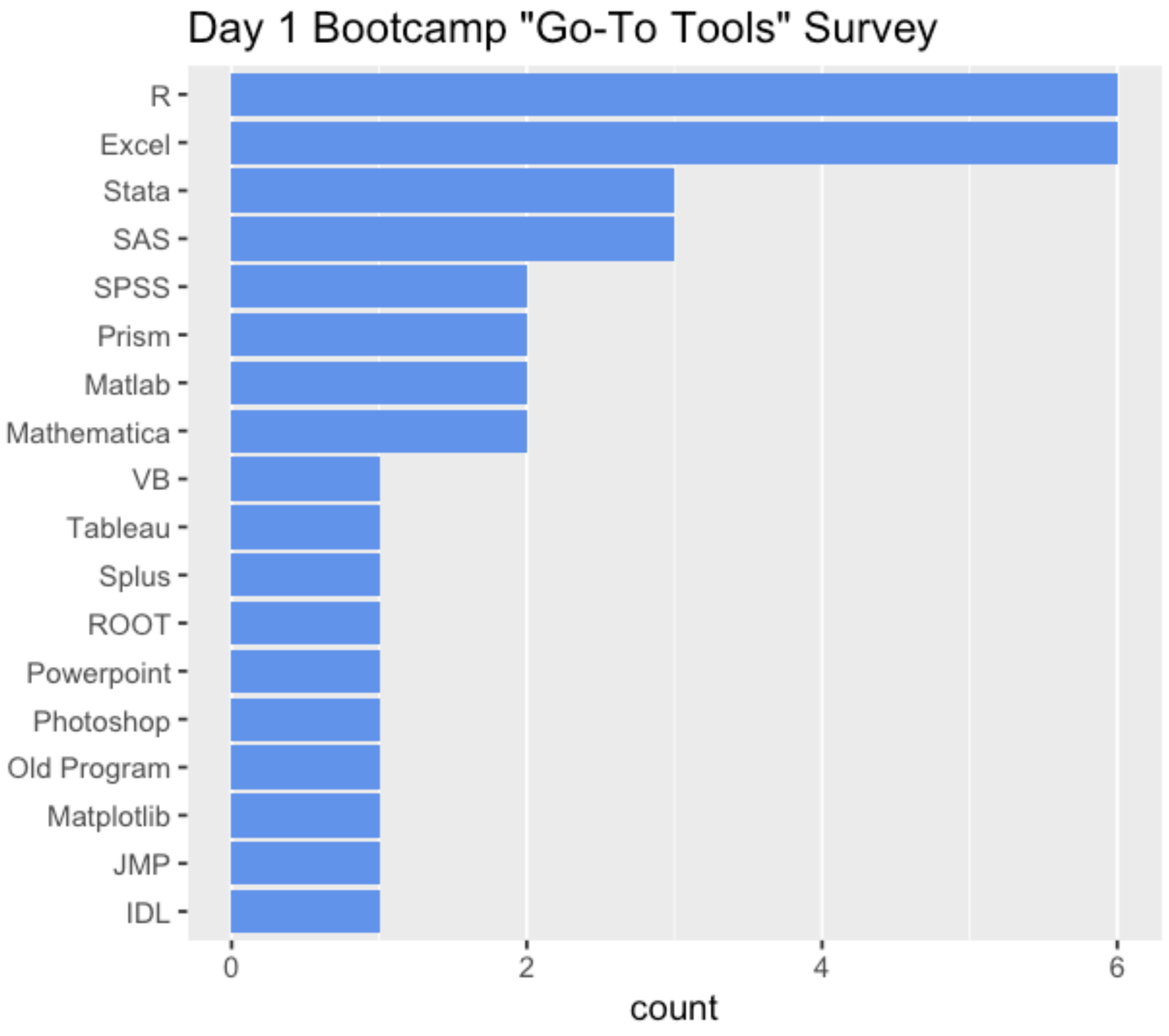
Misleading Graph



You-draw-it

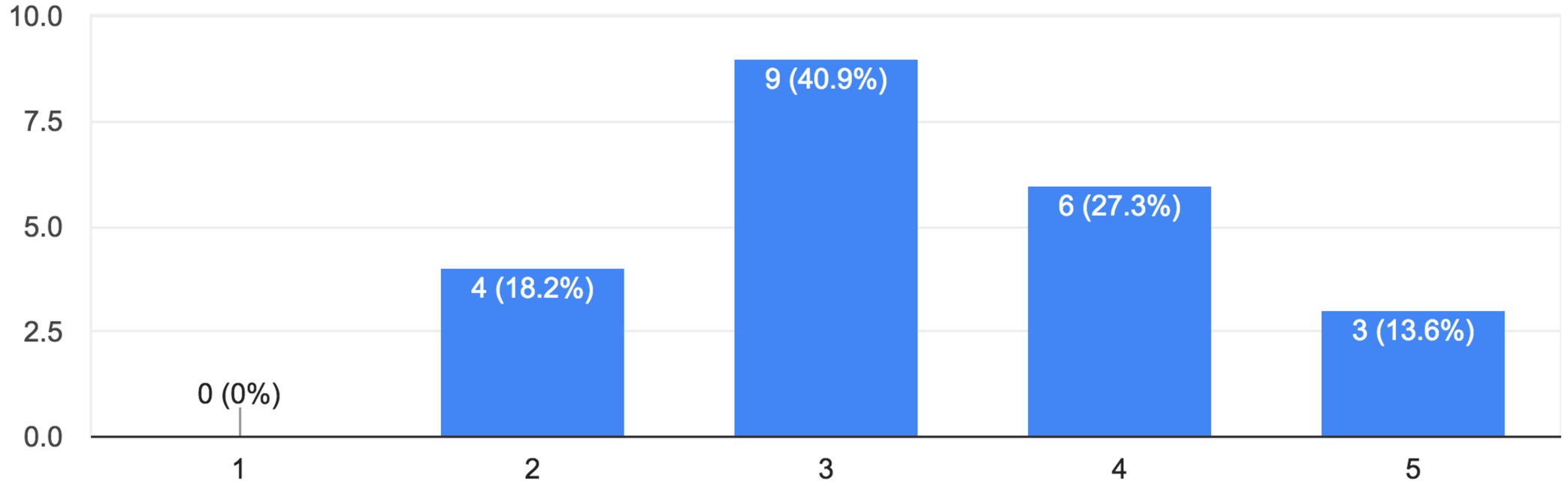

Under President Obama, the **unemployment rate** ...
Draw the line for the Obama years.
5.8%
4.7%
BUSH YEARS
OBAMA YEARS
2001
'04
'08
'12
'16
Show me how I did.

GRAPHING TOOLS


Day 1 Bootcamp "Go-To Tools" Survey
R
Excel
Stata
SAS
SPSS
Prism
Matlab
Mathematica
VB
Tableau
Splus
ROOT
Powerpoint
Photoshop
Old Program
Matplotlib
JMP
IDL
0
2
4
6
count

Please indicate your preference for workshop content on a scale from 1 to 5, where 1 = all data visualization theory and 5 = all tools (primarily ggplot2).

22 responses



Why R?

- virtually unlimited graphical options
- opinionated graphics
- analytical tools (> 13,000 CRAN packages)
- community
- reproducibility
- ease of workflow: everything in one document
- free and open source

Why not R?

- learning curve
- lack of GUI for graphics
- interactive graphics are not native

Graphics in R

pie {graphics} R Documentation

Pie Charts

Description

Draw a pie chart.

Usage

```
pie(x, labels = names(x), edges = 200, radius = 0.8,
    clockwise = FALSE, init.angle = if(clockwise) 90 else 0,
    density = NULL, angle = 45, col = NULL, border = NULL,
    lty = NULL, main = NULL, ...)
```

Arguments

`x`	a vector of non-negative numerical quantities. The values in `x` are displayed as the areas of pie slices.
`labels`	one or more expressions or character strings giving names for the slices. Other objects are coerced by `as.graphicsAnnot`. For empty or `NA` (after coercion to character) labels, no label nor pointing line is drawn.
`edges`	the circular outline of the pie is approximated by a polygon with this many edges.
`radius`	the pie is drawn centered in a square box whose sides range from *-1* to *1*. If the character strings labeling the slices are long it may be necessary to use a smaller radius.
`clockwise`	logical indicating if slices are drawn clockwise or counter clockwise (i.e., mathematically positive direction), the latter is default.

Sources

"John Snow, Cholera Map" https://www1.udel.edu/johnmack/frec682/cholera/

"Data Science Process" diagram: Hadley Wickham and Garrett Grolemund, R for Data Science, 1.1 r4ds.had.co.nz/introduction.html

"Growth of Data Visualization": Hadley Wickham, 2013, "Graphical Criticism: Some Historical Notes", p. 43 www.tandfonline.com/doi/full/10.1080/10618600.2012.761140

"Perception Studies", "Cleveland Dot Plot": Naomi Robbins, 2013, *Creating More Effective Graphs*, Ch. 1., Ch. 3

"Wrong / Misleading Graphs" http://www.mediaite.com/tv/fox-news-airs-seriously-misleading-obamacare-graphic/

"Florence Nightingale's Coxcomb Diagram, 1858" https://understandinguncertainty.org/coxcombs

"Nightingale's Data, Redrawn" Andrew Gelman and Antony Unwin, 2012. "Infovis and Statistical Graphics: Different Goals, Different Looks" http://www.stat.columbia.edu/~gelman/research/published/vis14.pdf