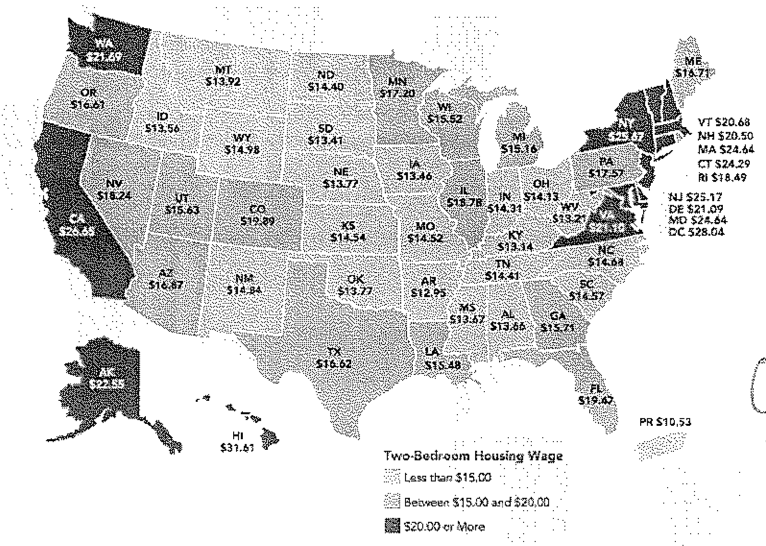
Today: Critiquing Statistical Graphics
Introduction to Data
Graphics Principles
Friday: Introduction to R and Reproducibility
Monday: No class (Martin Luther King, Jr. Day)

Sam Ventura
36-315

Department of Statistics
Carnegie Mellon University

January 13, 2016

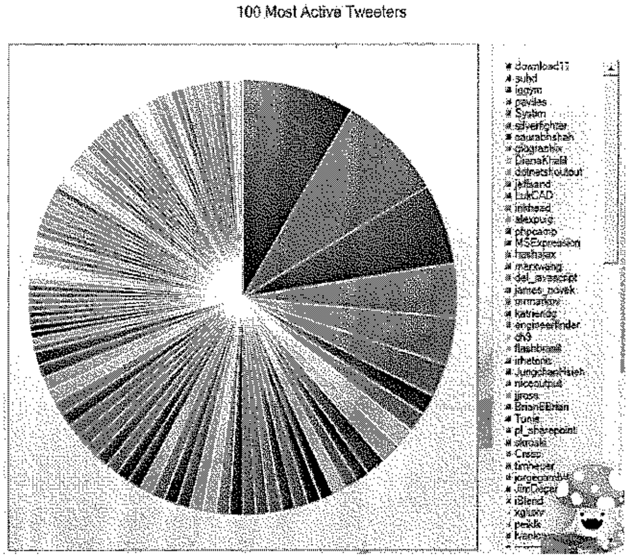
Hourly Wages to Afford Two-Bedroom Apartment



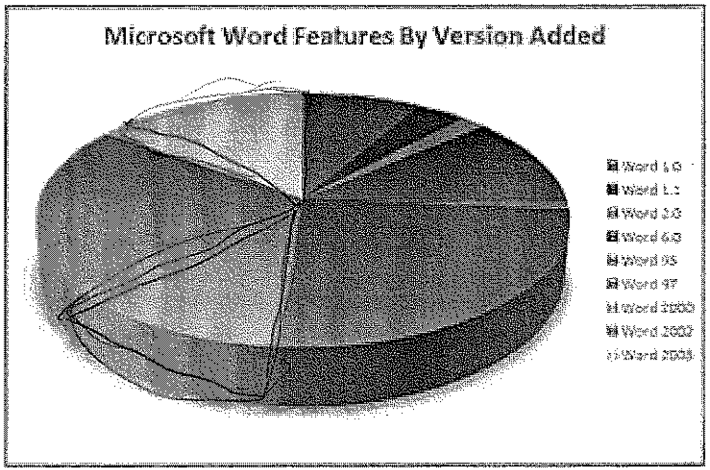
improve
w/ color
scale
OR just
more options

Top 100 Tweeters

Pie
=
bad



Never Make 3-D Pie Charts



distortion

What Is Data?

Information organized in some fixed easy-to-understand way

Ex) Tweets

field goal attempts $\left\{ \begin{array}{l} \text{made} \\ \text{missed} \end{array} \right.$ → 3pt sh %

Temperature

Censuses / Surveys → collect info on population, demographics, etc

How Do We Describe Data? mean, median, mode

Two measurements → rows used to describe datasets:

n = # of obs, people, subjects, objects, etc

p/d = # of covariates, variables, questions, etc → columns

Data is usually in matrix form:

$$\begin{array}{ccc} & & V_1 \quad V_2 \cdots \quad V_d \\ O_1 & X_1 & \left[\begin{array}{c} \square \\ \\ \\ \\ \\ \end{array} \right. \\ O_2 & X_2 & \\ O_3 & X_3 & \\ \vdots & \vdots & \\ O_n & X_n & \left. \right] \end{array}$$

rows → observations

~~eg ex~~

single row has all answers to all ?s from a single person

columns → variables

single column has all answers ~~for~~ from a single ? for all people

Types of Data → String, integer

Categorical → qualitative, describes qualities of obs

Ordered: strongly disagree, disagree, neutral, agree, SA

- educ. level

unordered: "nominal" → race, colors (sort of)

names / general text, gender

Continuous:

↳ real-valued, quantitative, numerical data

Notation $X = \{X_1, X_2, \dots, X_d\}$,

↓ data/variable

$X_i \in \mathbb{R}$

→ double, int, float

$X \in \mathbb{R}^d$

Graphics and Their Goals (from Tufte) → father of graphics

Graphics: visually display measured quantities by combining points, lines, coordinate system, numbers, symbols, words, shading, color

Goals: show data!

- induce viewer to think about substance, not graphical methodology
- avoid **distorting** the data
- present numbers in small space
- make large, complicated datasets more coherent
- encourage comparison of different pieces of data
- reveal data at several levels of detail
- describe, explore, tabulate, or decorate
- be closely integrated with statistical/verbal descriptions of dataset

Graphs that do not meet these goals are not successful

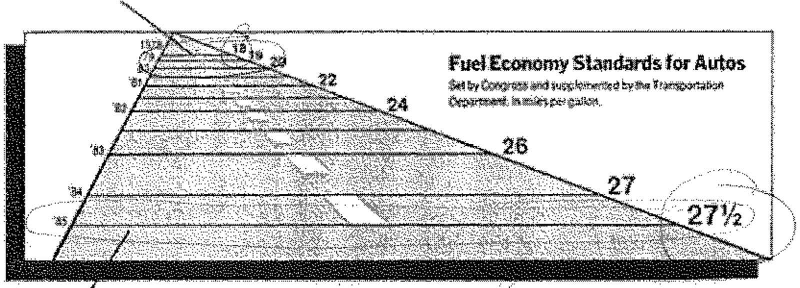
Graphs leading viewers to make misleading conclusions should be avoided

Distortion

Visual representation of data is inconsistent with numerical representation

In other words: **The graph doesn't match the data**

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

Optimal: $LF \approx 1$

$LF > 1$ → enhance the effect

$LF < 1$ → decrease the effect

Lie Factor

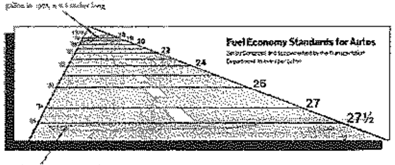
Tufte suggests optimizing the Lie Factor:

$$LF = \frac{\text{size of "effect" in graphic}}{\text{- - - - - - - in data}}$$

"effect" = change in amount of some feature or variable

Fuel Economy Standards Example:

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

→ Actual % increase (in data)

$$\frac{|27.5 - 18|}{18} \approx 0.528$$

graphical increase (in graph)

$$\frac{|5.3\text{ in} - 0.6\text{ in}|}{0.6\text{ in}} = 7.83$$

$$LF = \frac{7.83}{0.528} = 14.83$$

"Decorating" / Data-Ink

Graphics should not draw the viewer's attention away from the data.
Extras get in the way.

Note: Decoration does not refer to appropriate graph labeling.
Labels should always be clear, detailed, and thorough.
Label key parts of the data. Add text explanations if necessary.

Data Ink should primarily present information about the data:
the non-erasable, non-redundant core of a graphic

Tufte suggests using the *data-ink ratio*:

$$DI = \frac{\text{data ink}}{\text{total ink on graphic}}$$

% of ink devoted to non-redundant / useful information.

Ideally → Maximize DI (max = 1)
won't quite get to 1, because of axes, grid lines etc

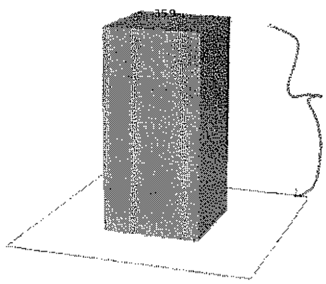
"Decorating" / Data-Ink

Two ways to increase the proportion of data-ink:

Remove non-data-ink:
↳ Ink that does not depict statistical info

In class wednesday 1/20, hands on map graphic

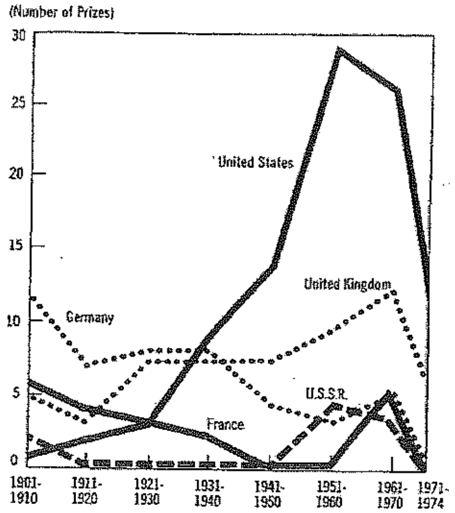
Remove redundant data-ink: → Ink that is unnecessarily redundant / repetitive

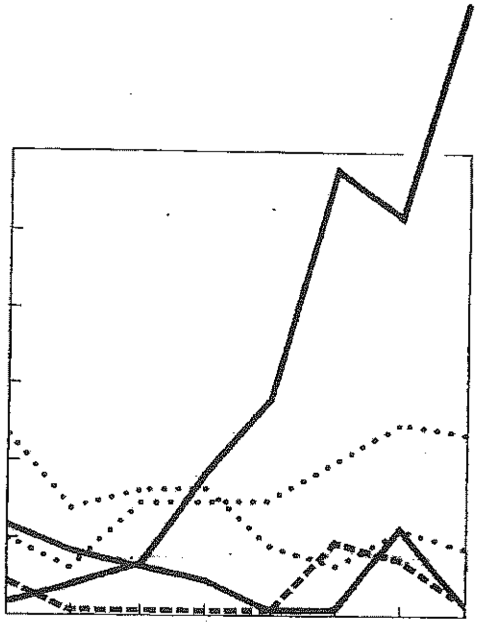


Indications of height:

1) height of front-left line on bar
2) height of front-right line on bar
3) - - - - back-right - - - -
4) position of number
5) value of number
6) Position of top line (front)
7) position of top line (back)
8) height of shading
9) etc....

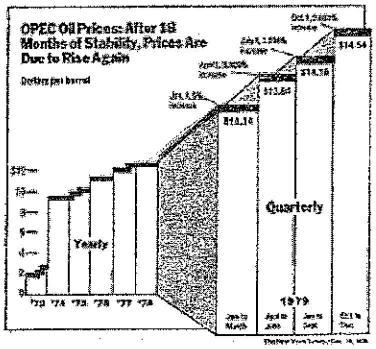
Nobel Prizes Awarded in Science, for Selected Countries, 1901-1974





Graph Principles (Tufte)

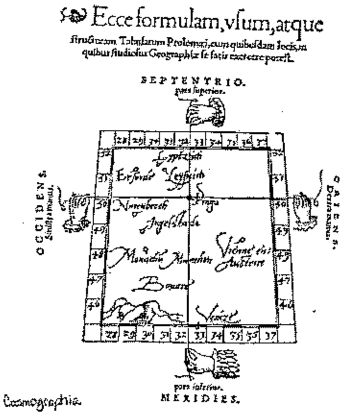
Minimize **Design Variation**:   Changes in Design may imply Changes in Data



New York Times, December 19, 1978, p. D-9.

Graph Principles (Tufte)

Maximize **Data-Ink Ratio:** Ink in graphic should primarily show data



1546 Cosmographia

Size and Types of Data

Two measurements used to describe data:

n # of obs (people, subjects, objects)

p/d # dim or var (variables or questions)

Data usually in matrix form:

$$\begin{array}{c|cccc} & V_1 & V_2 & \cdots & V_d \\ X_1\ O_1 & p_1 & p_2 & \cdots & p_d \\ X_2\ O_2 & & & & \\ \vdots & & & & \\ X_n\ O_n & & & & \end{array}$$

rows → observations
answers to all the questions for one obs

cols → variables/questions
answers for all obs to one question

In R: dim(data) = n×d

Often the # of colums is the important piece of the dim; how we graph/visualize depends on d (also to some extent on the n; more later)

Two major types we'll be working with: *categorical* and *contiuous*

Categorical: Qualitative, describes qualities of the obs string, int (factor?)

Non-ordered categ ex? can be numerical or text

- Favorite Ice Cream
- Nationality

Ordinal Categ: like having levels/factors ex?
Categ have an inherent order

e.g. Likert Scale: Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree

Continuous:

quantitative, numerical data int, double

often see notation like

$$X = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\} \in \mathbb{R}^d \leftarrow \text{real-valued number } d\text{-dim space}$$

$X \downarrow n \times d$ ↑ vector of d values, one for each var

Distortion

Visual representation of data is NOT consistent with the Numerical representation; *i.e. Graph doesn't match Data*

Graphics strongly dependent on visual perception of viewer.

Experiments have shown relationships between numerical measures and perceived measures. people look at different shapes, areas, lines, etc
& guess the length, area → big range of responses

<u>Area Example:</u>

Perceived Area (may) grow more slowly than Actual area

$PA = AA^{x}$ $x = 0.8 \pm 0.3$

Can't design graph for each viewer; What should we do?

some say Tables for 20 #'s or less → big debate, post discussion
well-labeled graph is fine as well; graphs for large sets & higher dim

Lie Factor

Tufte suggests optimizing the Lie Factor:

$$\text{Lie Factor} = \frac{\text{size of effect in graphic}}{\text{size of effect in data}}$$

effect - change/feature of interest

Optimal = 1 log LF = 0

LF < 1 graphic diminishes the effect < 0

LF > 1 graphic enhances the effect > 0
(more common)

Fuel Economy Standards Example: U.S. Congress and Department of Transportation set a series of fuel economy standards to be met by automobile manufacturers.

Actual % Increase $\frac{27.5 - 18}{18} = 0.528$

Graphical Increase $\frac{5.3 - 0.6}{0.6} = 7.83$

$LF = \frac{7.83}{0.528} = 14.83$

“Decorating” / Data Ink

Graphics should not draw the viewer’s attention away from the data. Extras get in the way.

Note: Decoration does not refer to appropriate graph labeling. Labels should always be clear, detailed, and thorough. Label key parts of the data. Add text explanations if necessary. More later.

Data Ink should be primarily present information about the data: the Non-Erasable core of a graphic, Non-redundant Ink.

Tufte suggests using the *data-ink ratio*:

DI: $\frac{\text{data ink}}{\text{total ink ingraphic}}$ = % ink devoted to non-redundant display of info about data

1 - % of graphic that could be erased

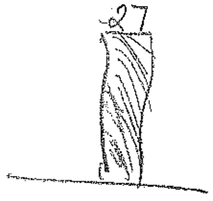
We want to maximize this ratio (within reason).

max of 1; 1 not realistic; why?
grid, border, etc

Two ways we can increase the proportion of data-ink:

Remove non-data-ink: ink that doesn't depict statistical info
see Playfair example; ~~re~~ disembodied hands

Remove redundant data-ink: Six indications of height



1) ht of left line
2) ht of right line
3) height of shading
4) position of top line
5) position of # 6) actual #

Removals should be done within reason; some redundancy will remain.