

Graphics in R

Part I: High-level graphics functions

We'll be working in this section with many of R's built-in data sets. To see a list of them, just type

```
> data()  
Data sets in package 'datasets':
```

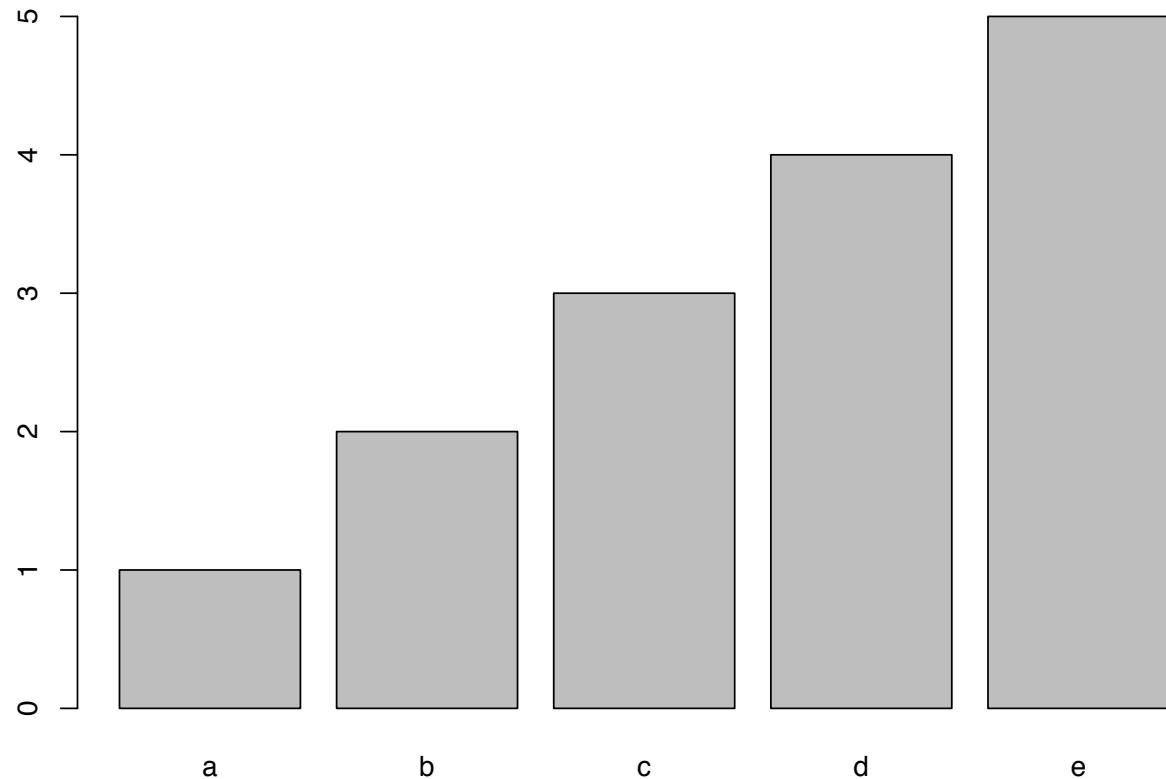
AirPassengers	Monthly Airline Passenger Numbers 1949-1960
BJsales	Sales Data with Leading Indicator
BJsales.lead (BJsales)	Sales Data with Leading Indicator
BOD	Biochemical Oxygen Demand
CO2	Carbon Dioxide uptake in grass plants
ChickWeight	Weight versus age of chicks different diets

...many more

Also, `help(AirPassengers)` and so on will give you information about each one.

I. Barplots

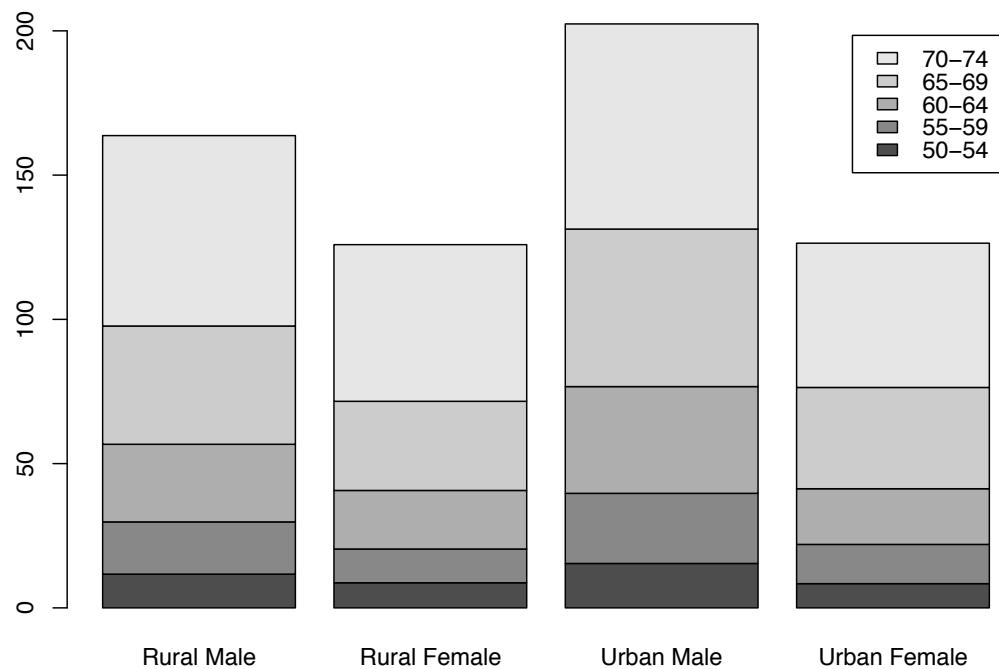
```
> x <- 1:5; names(x) <- letters[1:5]  
> x  
a b c d e  
1 2 3 4 5  
> barplot(x)
```



```
> VADeaths
```

	Rural Male	Rural Female	Urban Male	Urban Female
50-54	11.7	8.7	15.4	8.4
55-59	18.1	11.7	24.3	13.6
60-64	26.9	20.3	37.0	19.3
65-69	41.0	30.9	54.6	35.1
70-74	66.0	54.3	71.1	50.0

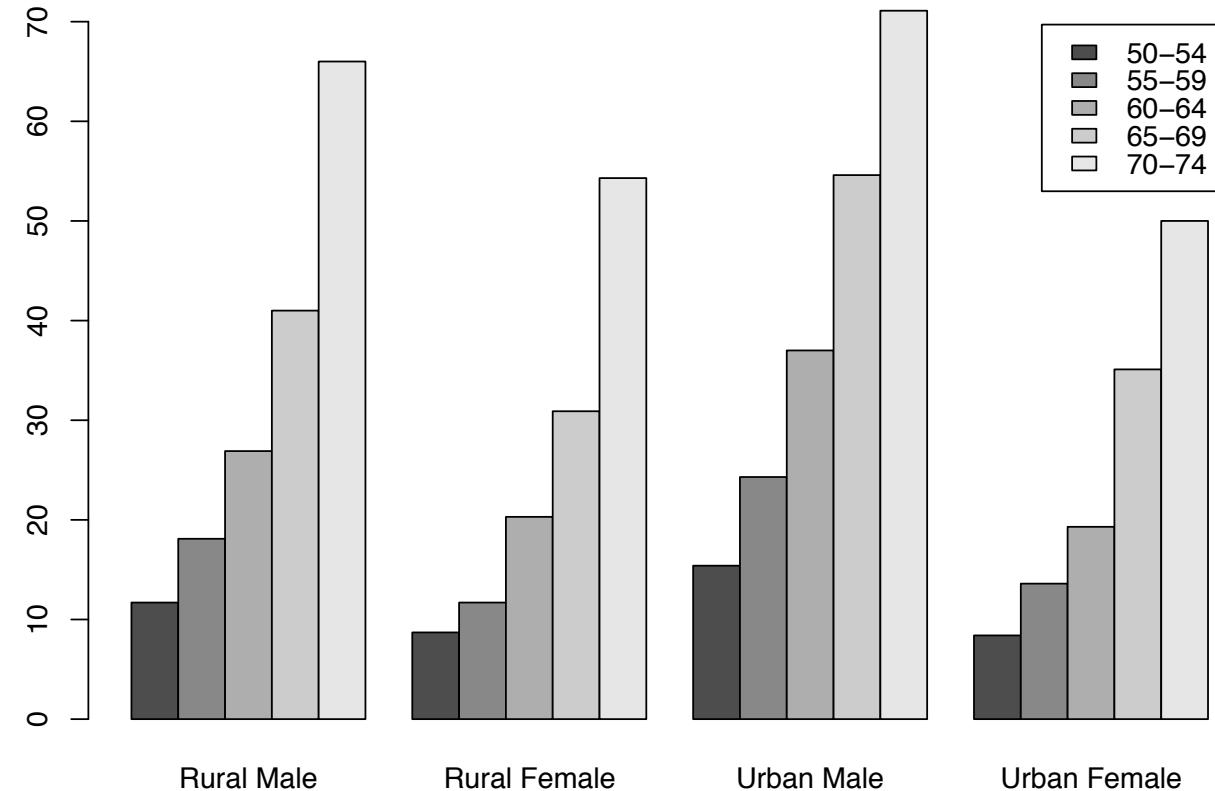
```
> barplot(VADeaths, legend = TRUE)
```



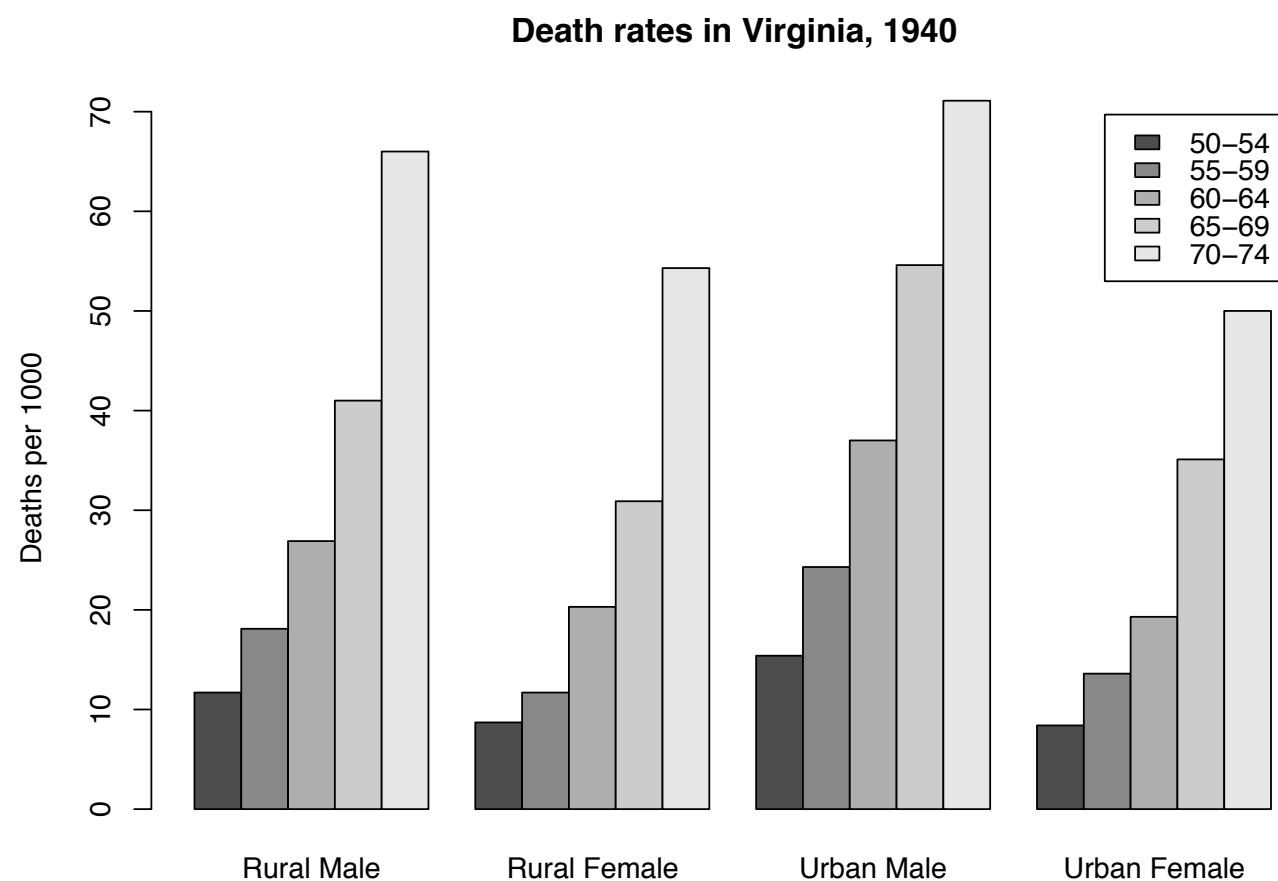
This stacked barplot makes it hard to read anything but the bottom category and the total.

Making a good plot in R is often a matter of iterative improvement.

```
> barplot(VADeaths, beside = TRUE, legend = TRUE)
```



```
> barplot(VADeaths, beside = TRUE, legend = TRUE,  
+         ylab = "Deaths per 1000",  
+         main = "Death rates in Virginia, 1940")
```



Aside: Saving your plots as graphics files

If you call a high-level plot command, R will automatically start a graphics device or window.

To save the contents of the already open device to a file, use `dev.print`.

```
> barplot(VADeaths, legend = TRUE)
> dev.print(device = pdf, file = "mybar.pdf",
+           height = 5, width = 6)          # Inches
> dev.print(device = jpeg, file = "mybar.jpeg",
+           height = 500, width = 600)      # Pixels
```

See `help(device)` for a list of other graphics formats.

Alternatively, if you are using the R GUI, you can click on the plot to highlight that window, then go to File > Save to save the image as a pdf file. However, it's nice to use the code version, so you can easily rerun your whole analysis, including generating the plots.

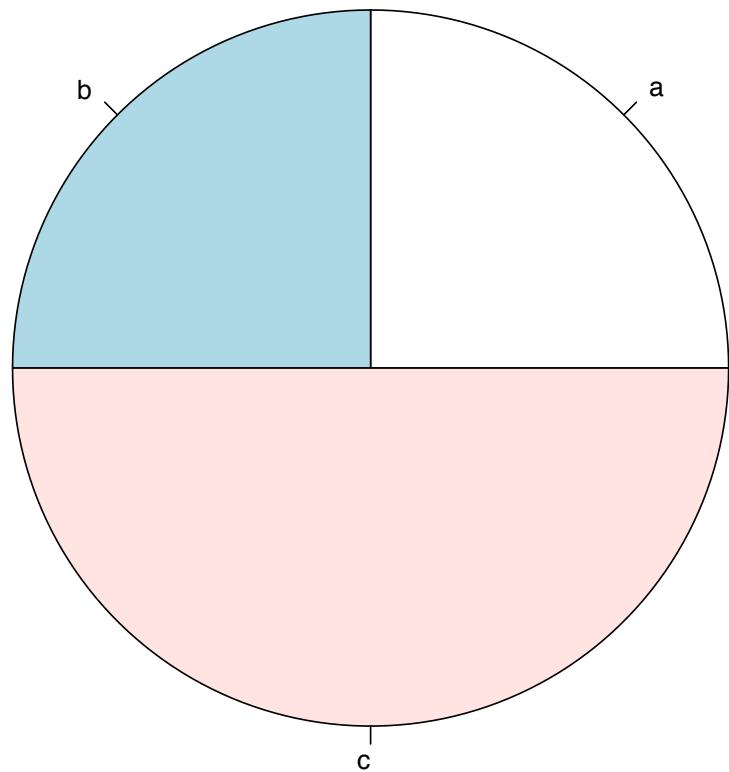
To close the device (shut the window), type

```
> dev.off()
```

To open multiple devices at once, use `x11()` or click on the little graphics icon in the GUI.

2. Pie charts

```
> pie(c(1, 1, 2), labels = letters[1:3])
```



Note that elements of the vector are normalized by their sum, so that the total gives 100% of the pie.

```
> Titanic  
, , Age = Child, Survived = No  
      Sex
```

Class	Male	Female
-------	------	--------

1st	0	0
2nd	0	0
3rd	35	17
Crew	0	0

```
, , Age = Adult, Survived = No  
      Sex
```

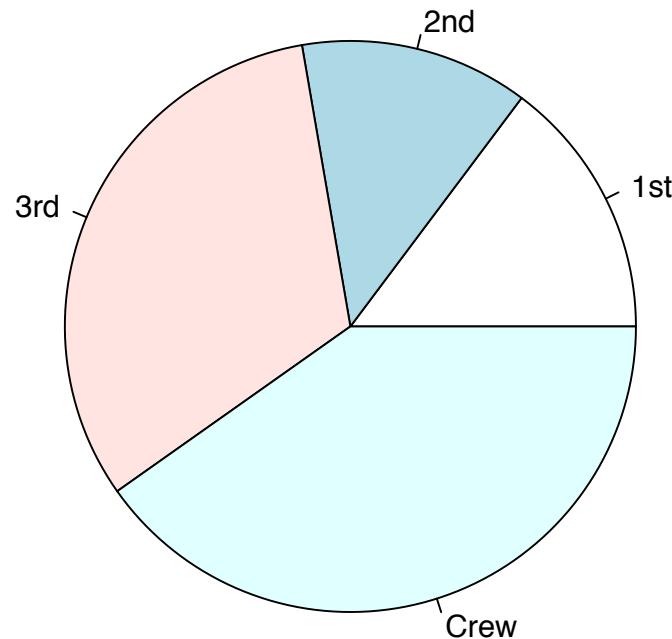
Class	Male	Female
-------	------	--------

1st	118	4
2nd	154	13
3rd	387	89
Crew	670	3

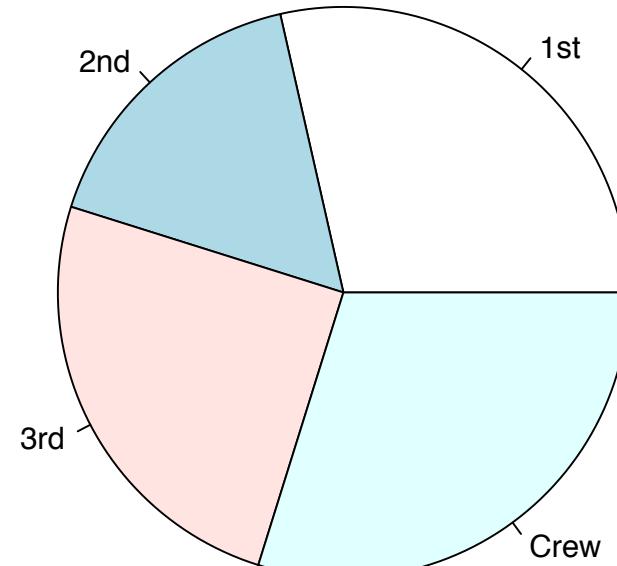
... two more matrices not printed here, with survivors
Did all groups have an equal survival rate?

```
> apply(Titanic, 1, sum) # Total passengers, each class  
1st 2nd 3rd Crew  
325 285 706 885  
> pie(apply(Titanic, 1, sum), main = "Total Passengers")  
> pie(apply(Titanic[,,, "Yes"], 1, sum),  
+      main = "Survivors")
```

Total Passengers



Survivors

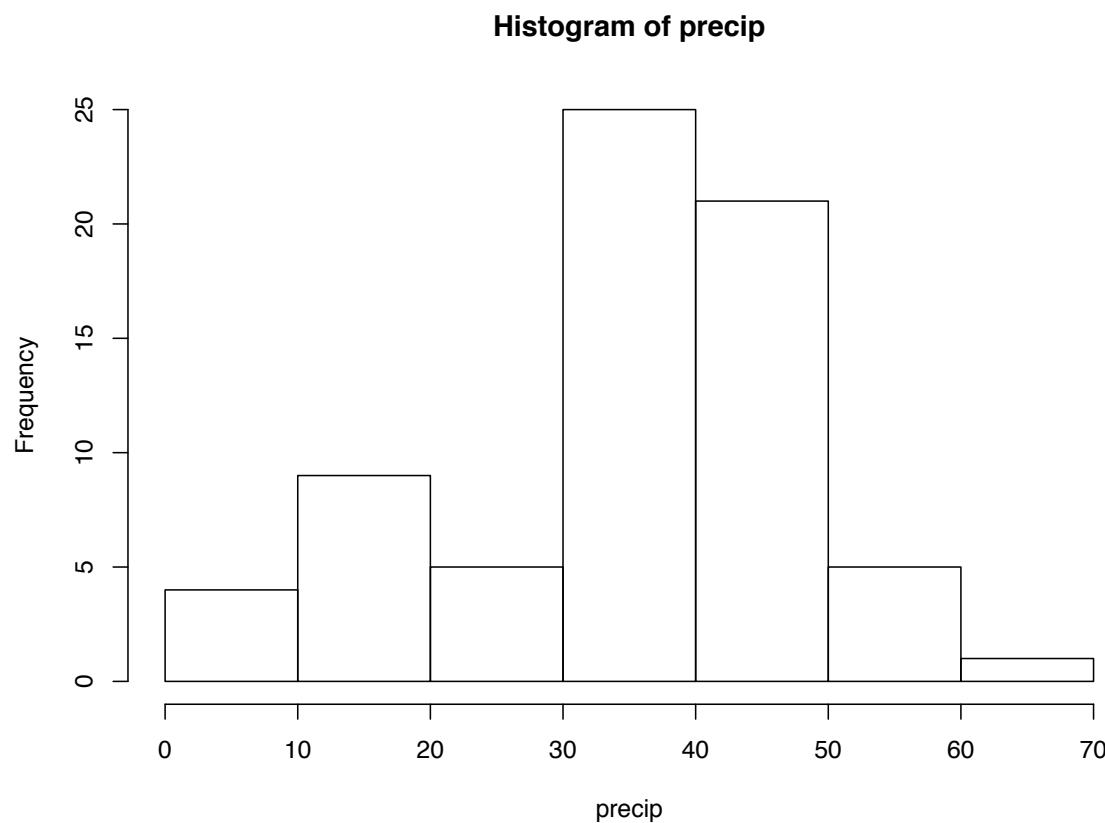


Studies of human perception show we are not very good at comparing areas, volumes, or angles.

- When making bar plots, start the axis at zero and keep all bars the same width, so that length and area are proportional.
- Try to avoid pie charts for anything requiring a precise comparison.

3. Histograms

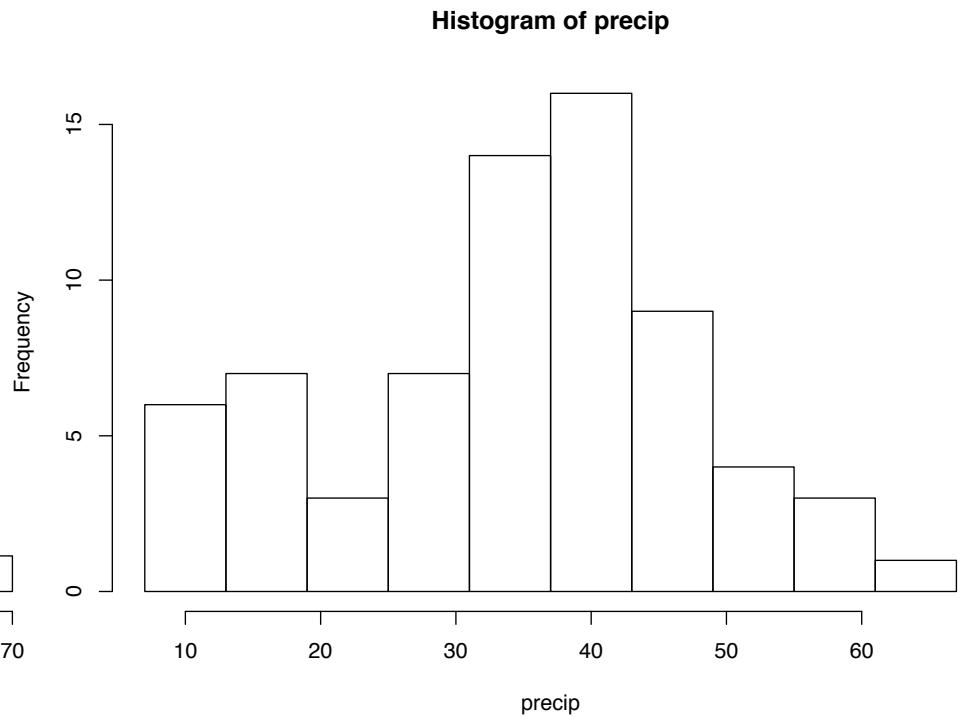
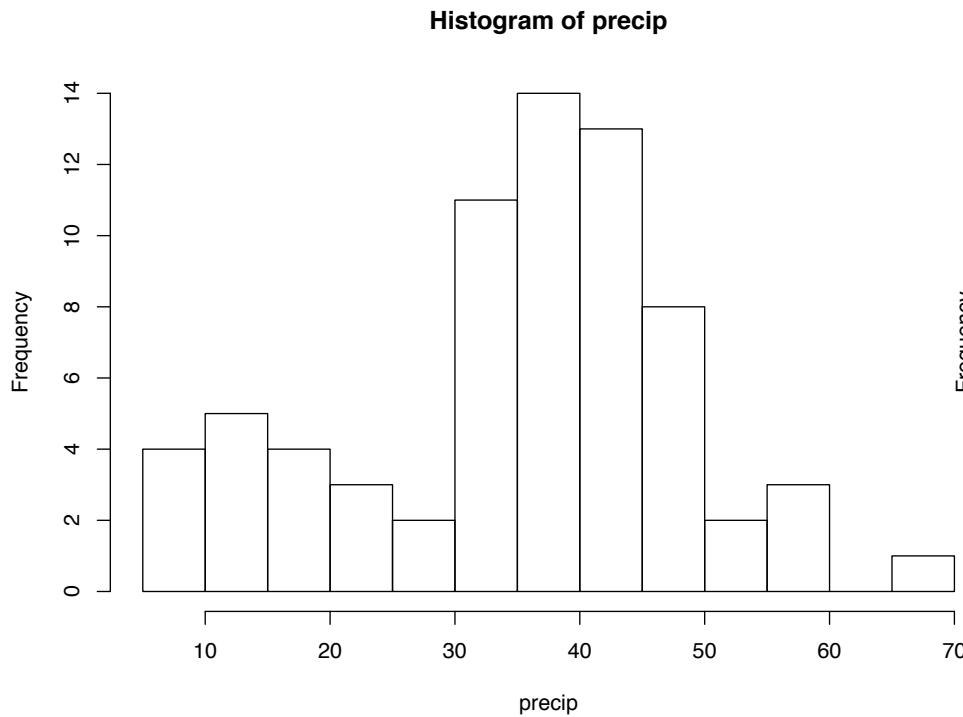
```
> precip[1:4] # Average annual precipitation in cities  
Mobile       Juneau      Phoenix Little Rock  
   67.0        54.7        7.0        48.5  
> hist(precip)
```



The height of the bars shows the number of observations falling into each bin.

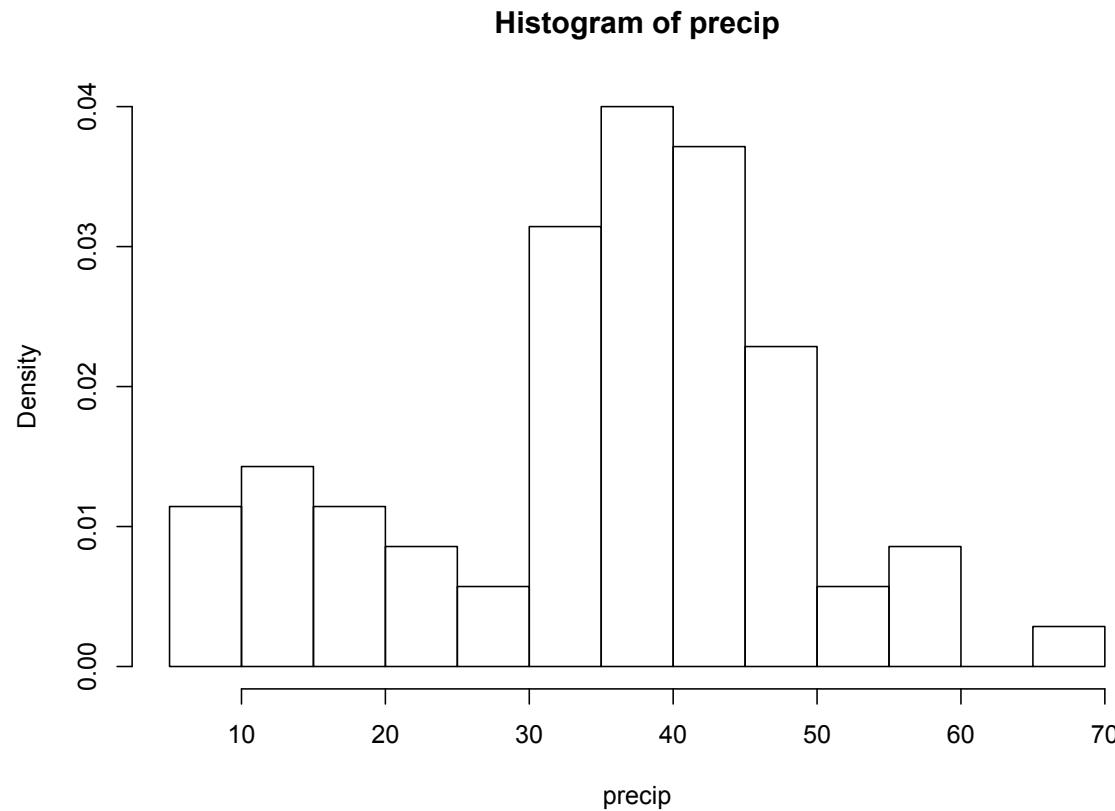
There are several ways to change the cutoff points.

```
> hist(precip, breaks = 10) # Only a suggestion to R  
> hist(precip, breaks = seq(min(precip), max(precip),  
+ length = 11)) # Force it
```



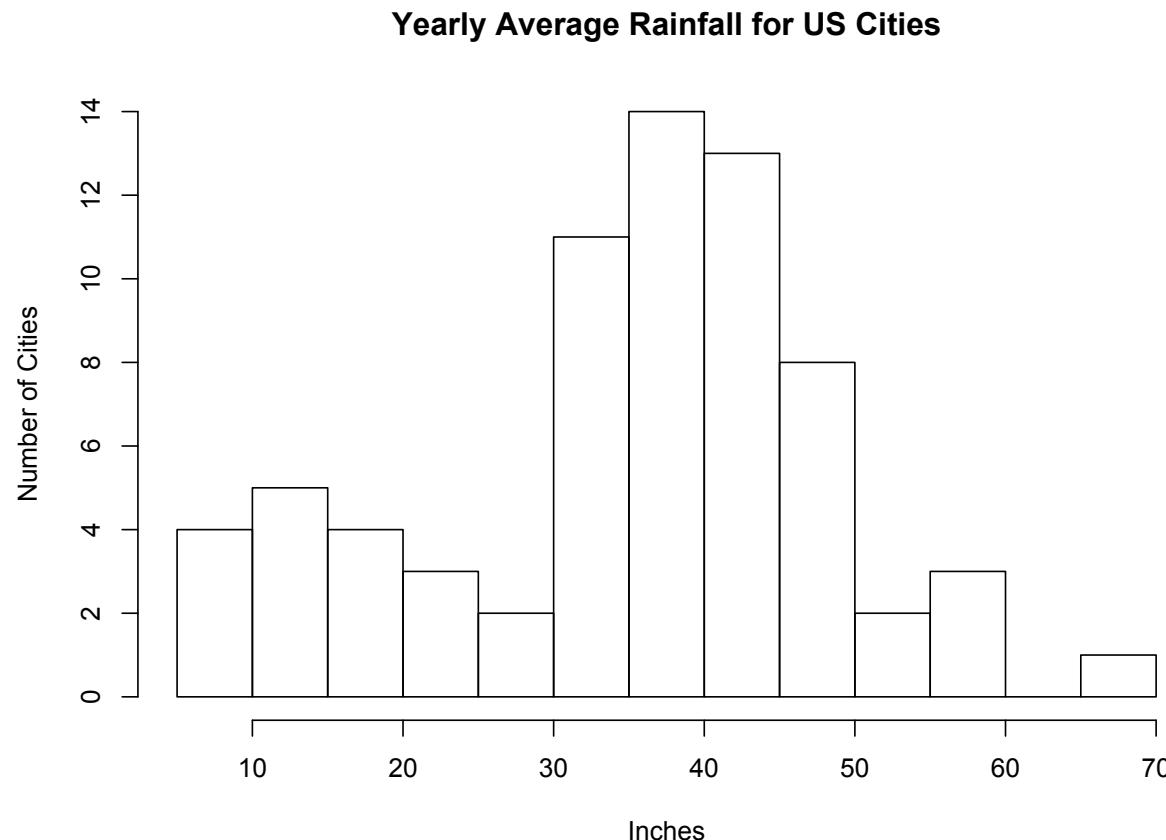
We can also change the y-axis to reflect probabilities rather than densities. Under the rescaled axis, the *areas of the bars will sum to 1*.

```
> hist(precip, breaks = 10, freq = FALSE)
```



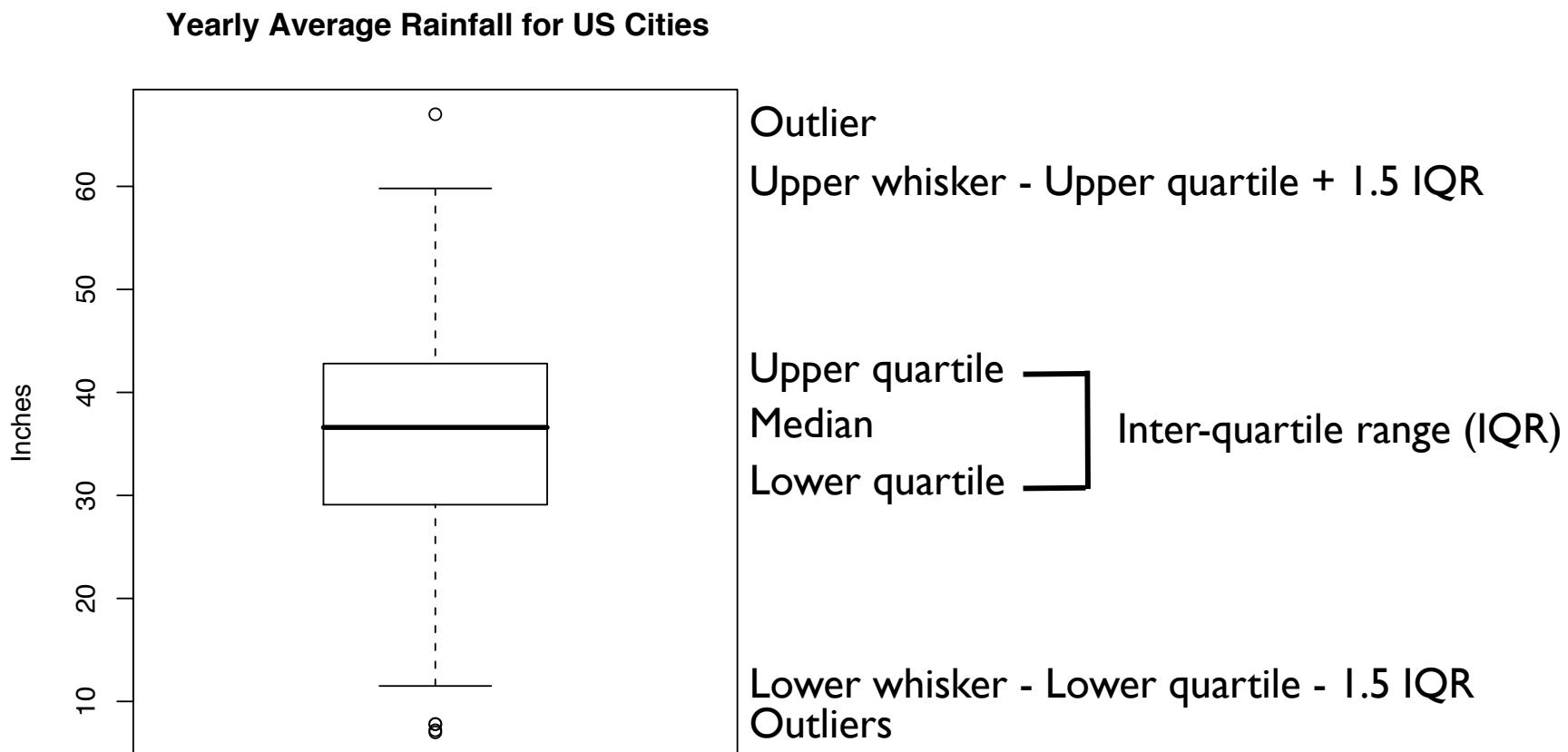
Again, let's add meaningful axis labels and a title.

```
> hist(precip, breaks = 10, xlab = "Inches",
+       ylab = "Number of Cities",
+       main = "Yearly Average Rainfall for US Cities")
```

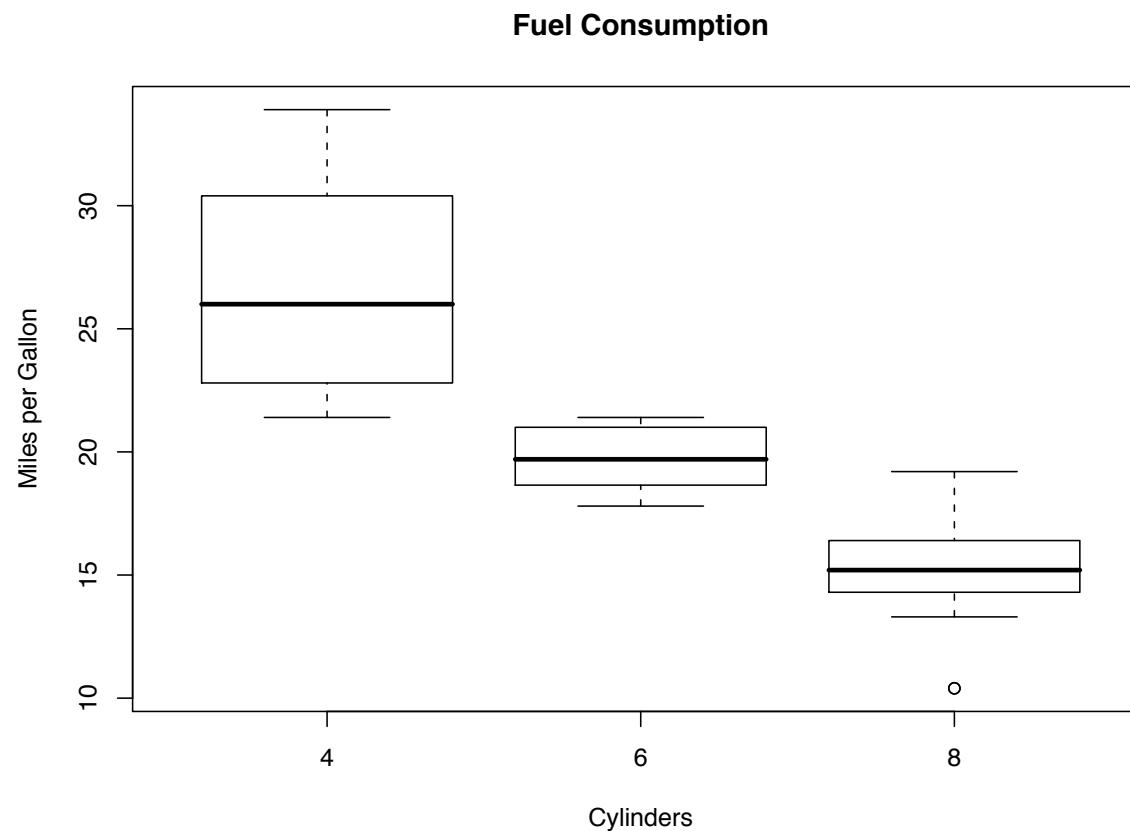


4. Boxplots

```
> boxplot(precip, ylab = "Inches",  
+           main = "Yearly Average Rainfall for US Cities")
```



```
> mtcars[1:2,1:5]
      mpg cyl disp hp drat
Mazda RX4     21   6 160 110 3.9
Mazda RX4 Wag 21   6 160 110 3.9
> boxplot(mpg~cyl, data = mtcars, xlab = "Cylinders",
+           ylab = "Miles per Gallon",
+           main = "Fuel Consumption")
```

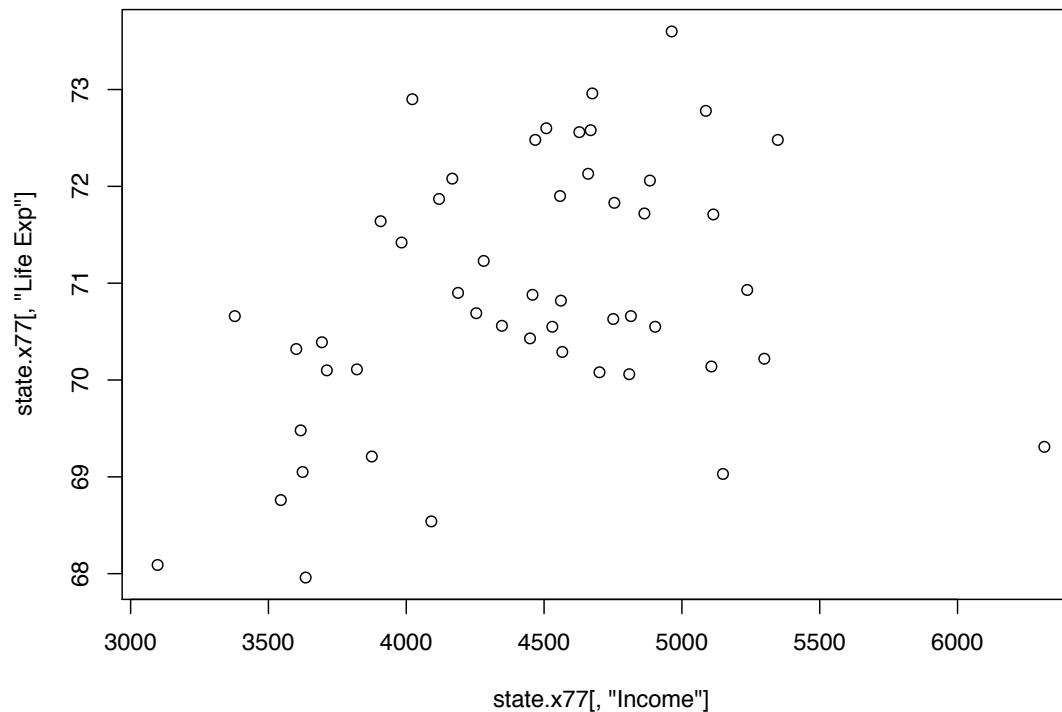


5. Scatterplots

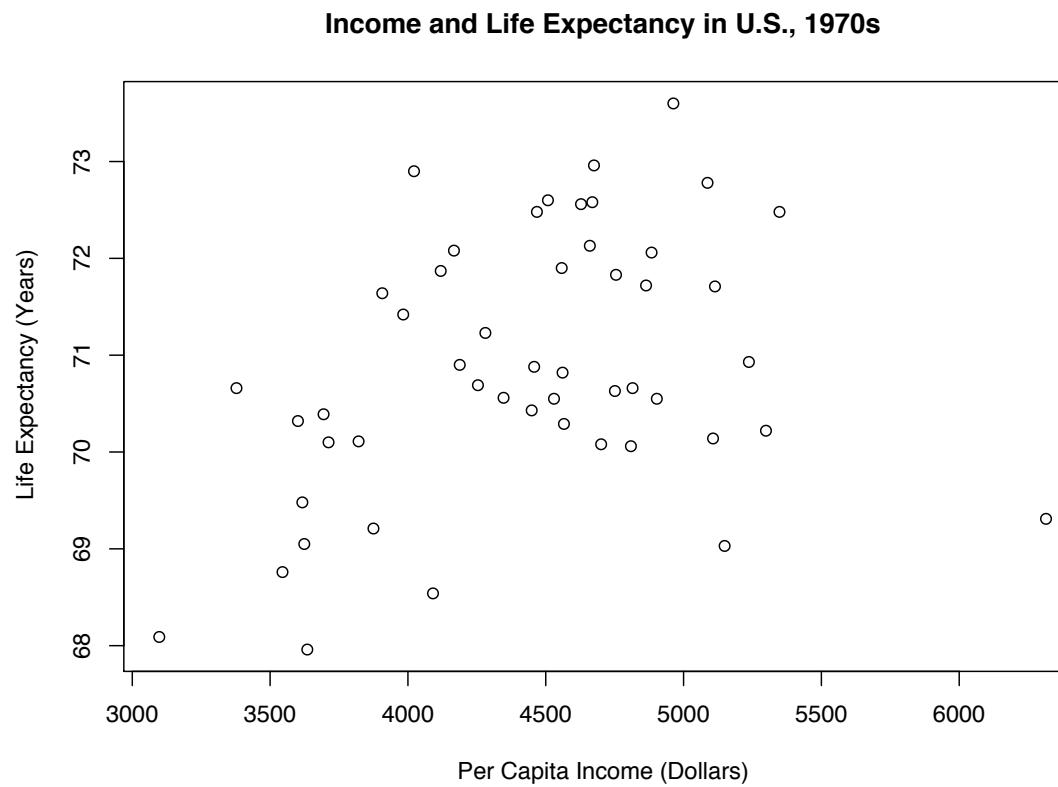
```
> state.x77[1:2,1:4]
```

	Population	Income	Illiteracy	Life Exp
Alabama	3615	3624	2.1	69.05
Alaska	365	6315	1.5	69.31

```
> plot(state.x77[, "Income"], state.x77[, "Life Exp"])
```



```
> plot(state.x77[, "Income"], state.x77[, "Life Exp"],  
+       xlab = "Per Capita Income (Dollars)",  
+       ylab = "Life Expectancy (Years)",  
+       main = "Income and Life Expectancy in U.S., 1970s")
```



We can label the interesting cases with `identify`.

Designing Good Graphics

(or, avoiding “The Dirty Dozen”)

Inspired by Wainer, H. (1984) “How to Display Data Badly.” *The American Statistician*, 38, 137-147.
Some additional images from Tufte, E. *The Visual Display of Quantitative Information* and online news sources.

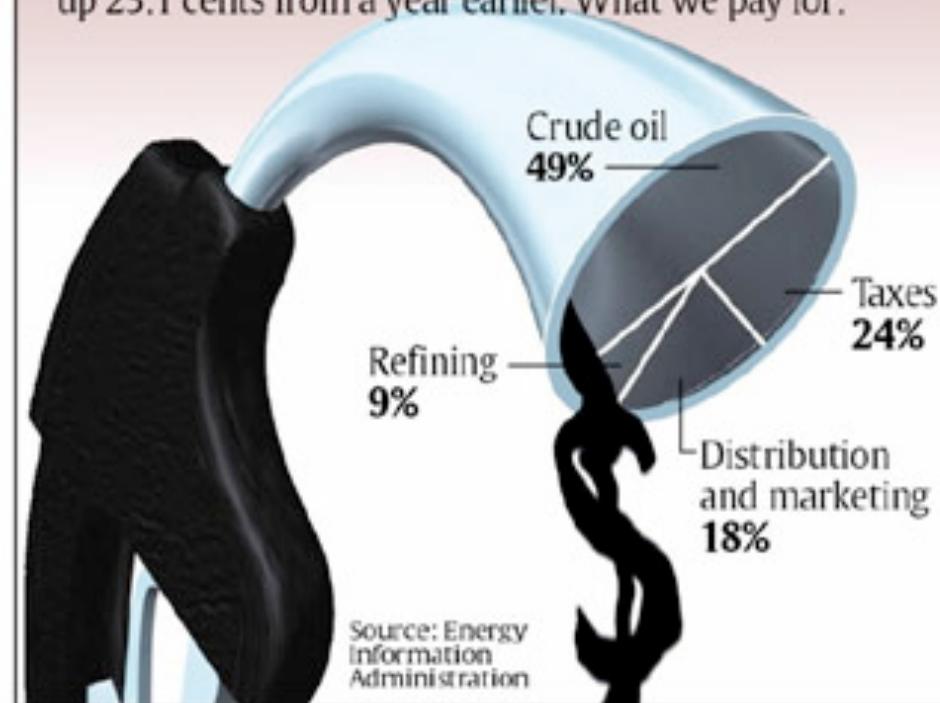
I. Show as few data as possible.

An example with lots of “chart junk,” not to mention visual distortion

USA TODAY Snapshots®

What gas bucks buy

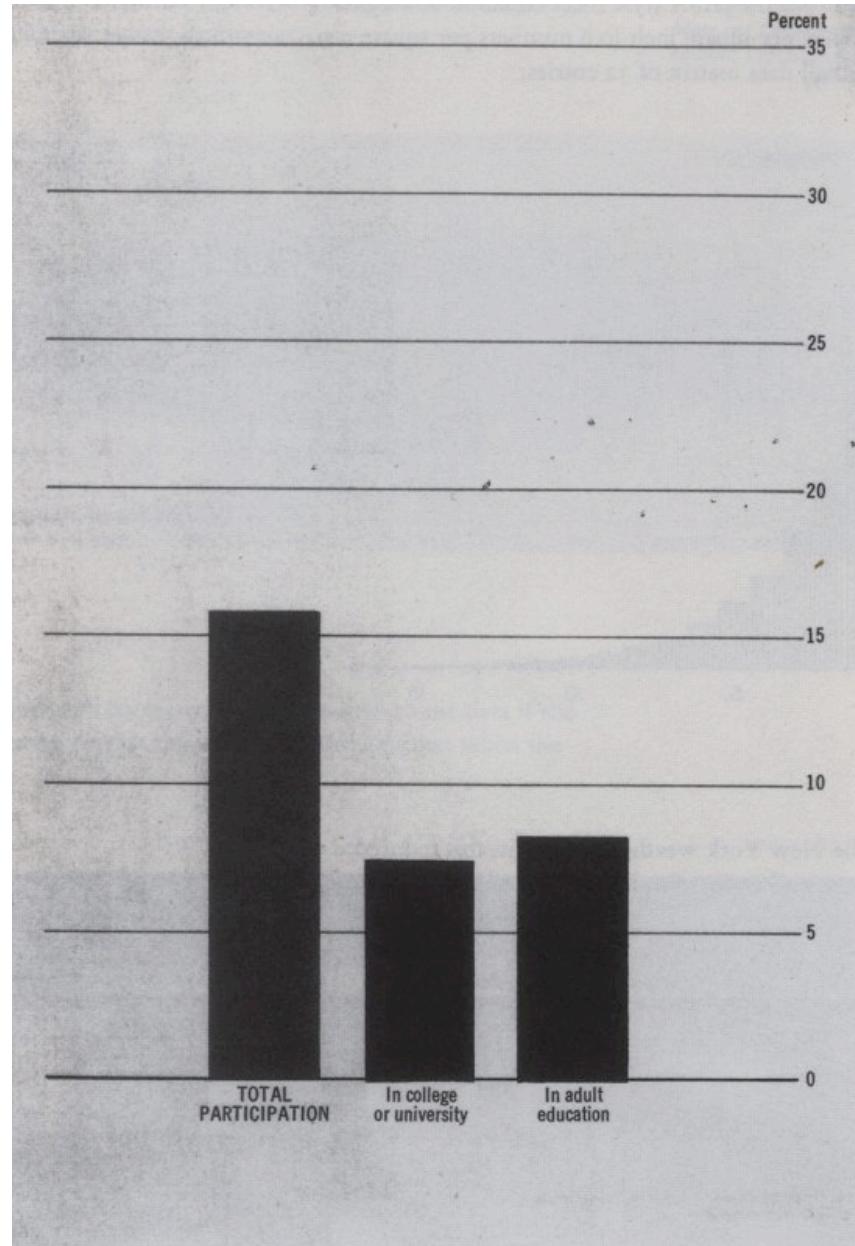
The average price of regular gas climbed to \$1.85 per gallon as of Jan. 24, up 3.4 cents from a week earlier and up 23.1 cents from a year earlier. What we pay for:



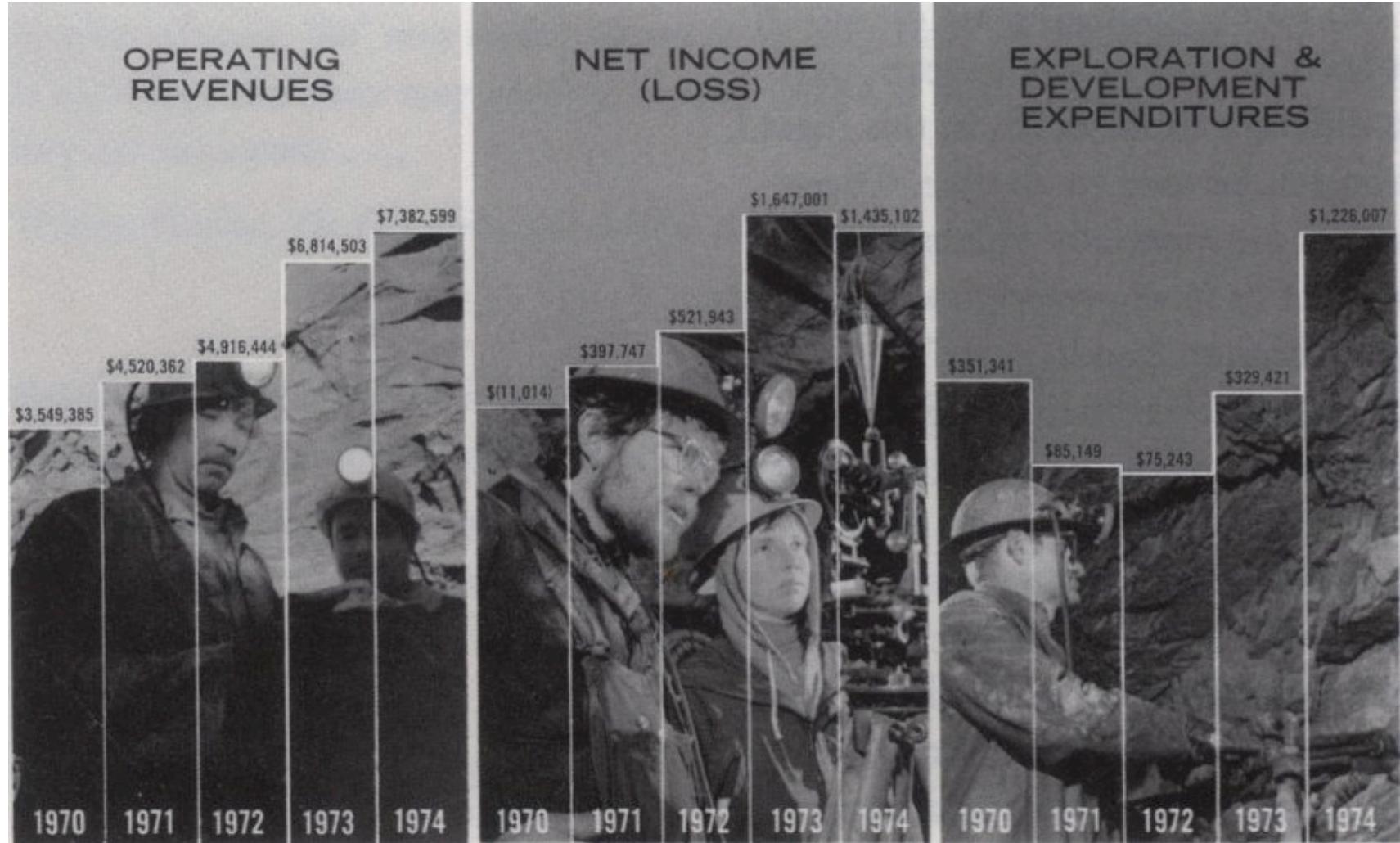
By Shannon Reilly and Marcy E. Mullins, USA TODAY

How many data points?

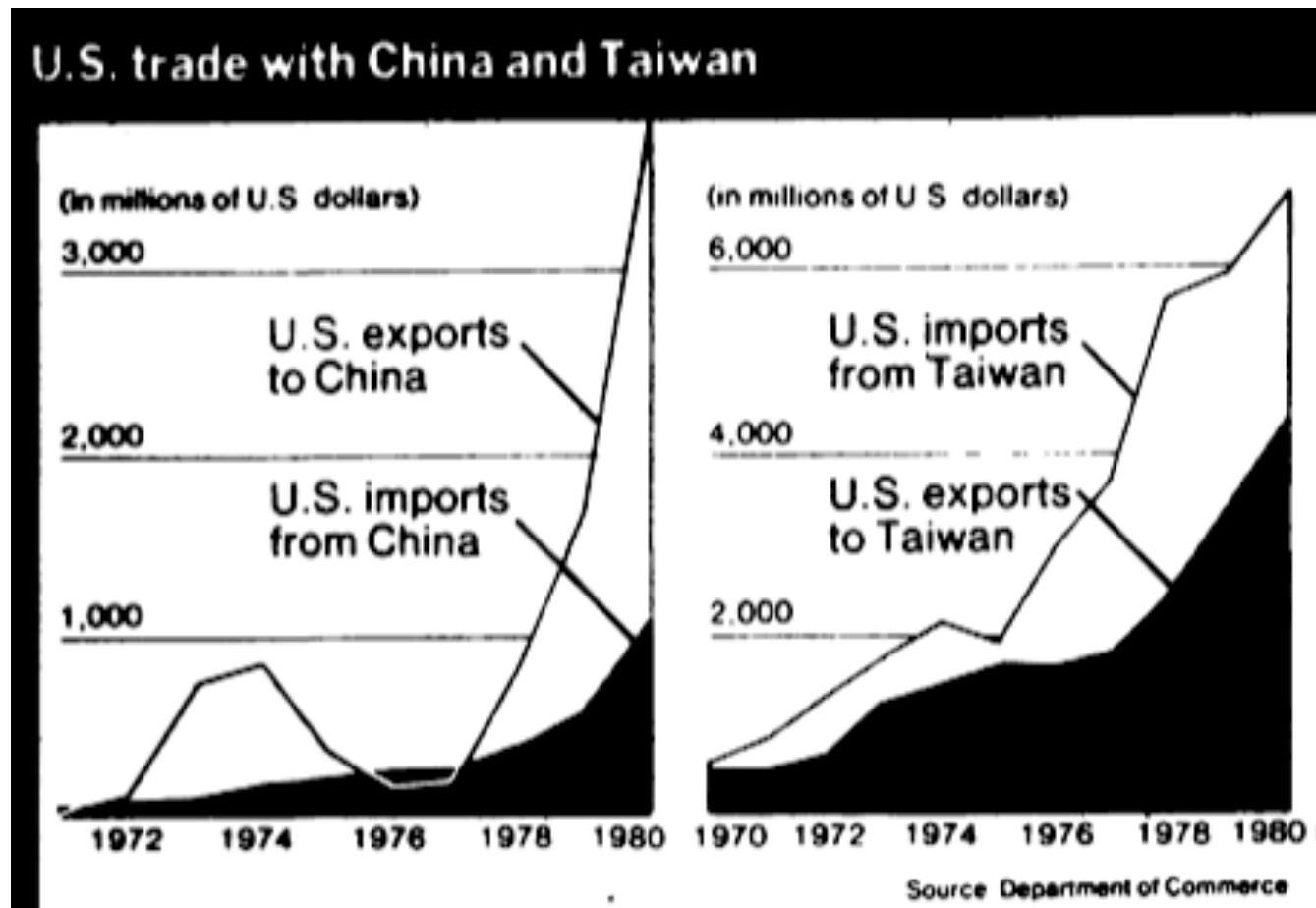
Tufte's “data to ink ratio”



2. Hide what data you do show.

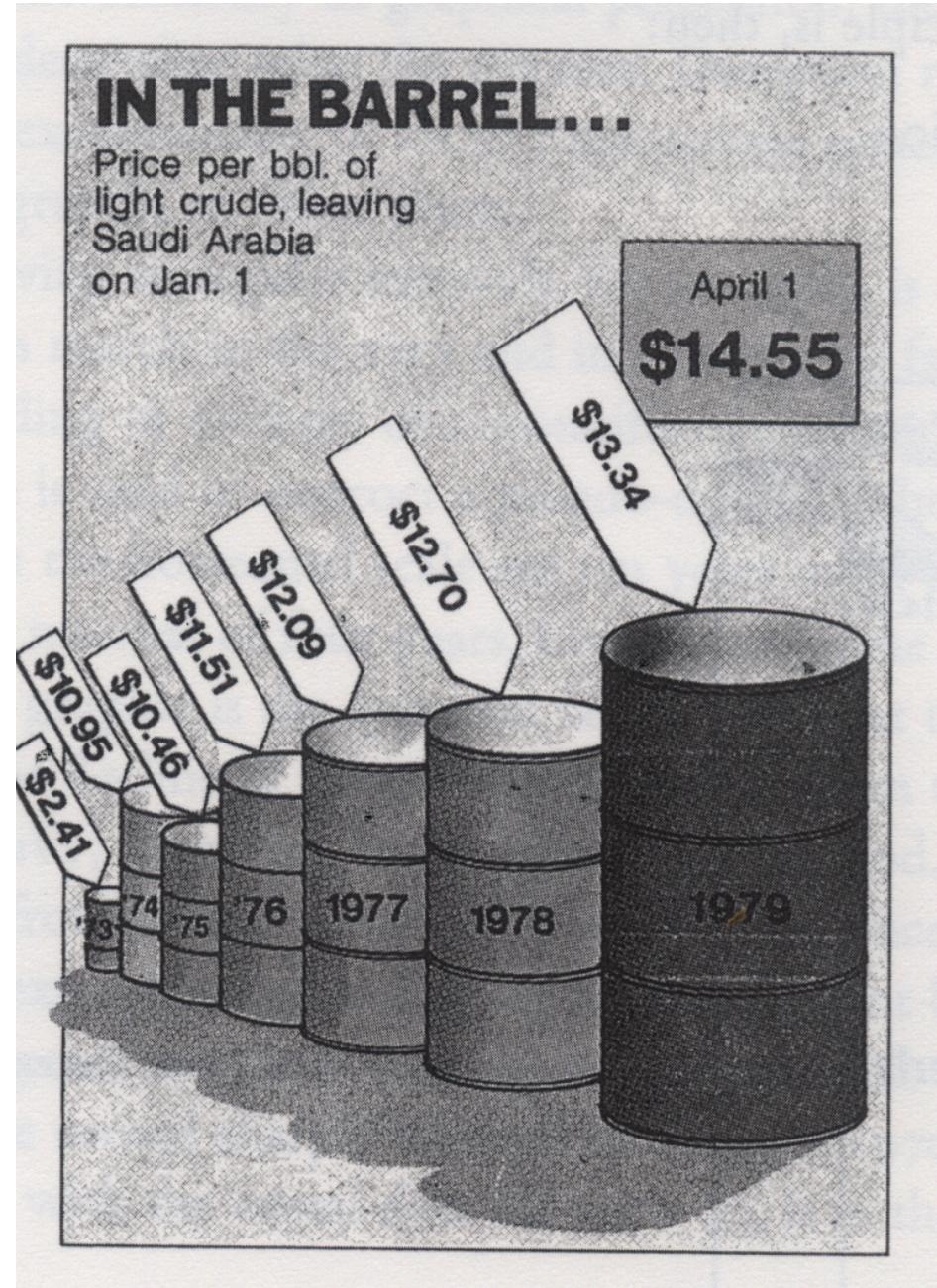


3. Ignore the visual metaphor, or reverse it mid-graph.



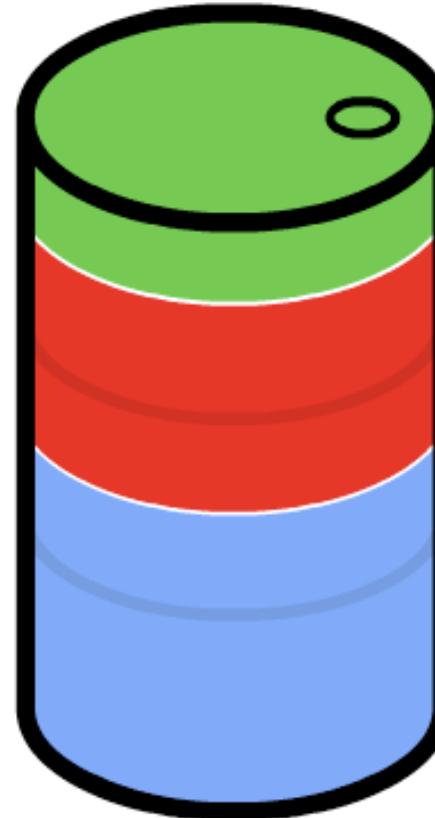
4. Only order matters.

Are we supposed to compare length, area, or volume?



Black gold at an all time high

Every \$1 increase in the price of a barrel of crude equates to roughly 0.42p on a litre of petrol at the pump. At \$150 a barrel, a litre of unleaded would cost £1.25 on the forecourt; at \$200 that rises to £1.46 - or £80 to fill up the family runabout.



Forecourt and transport

8.9%

Oil and refining

32.4%

Tax

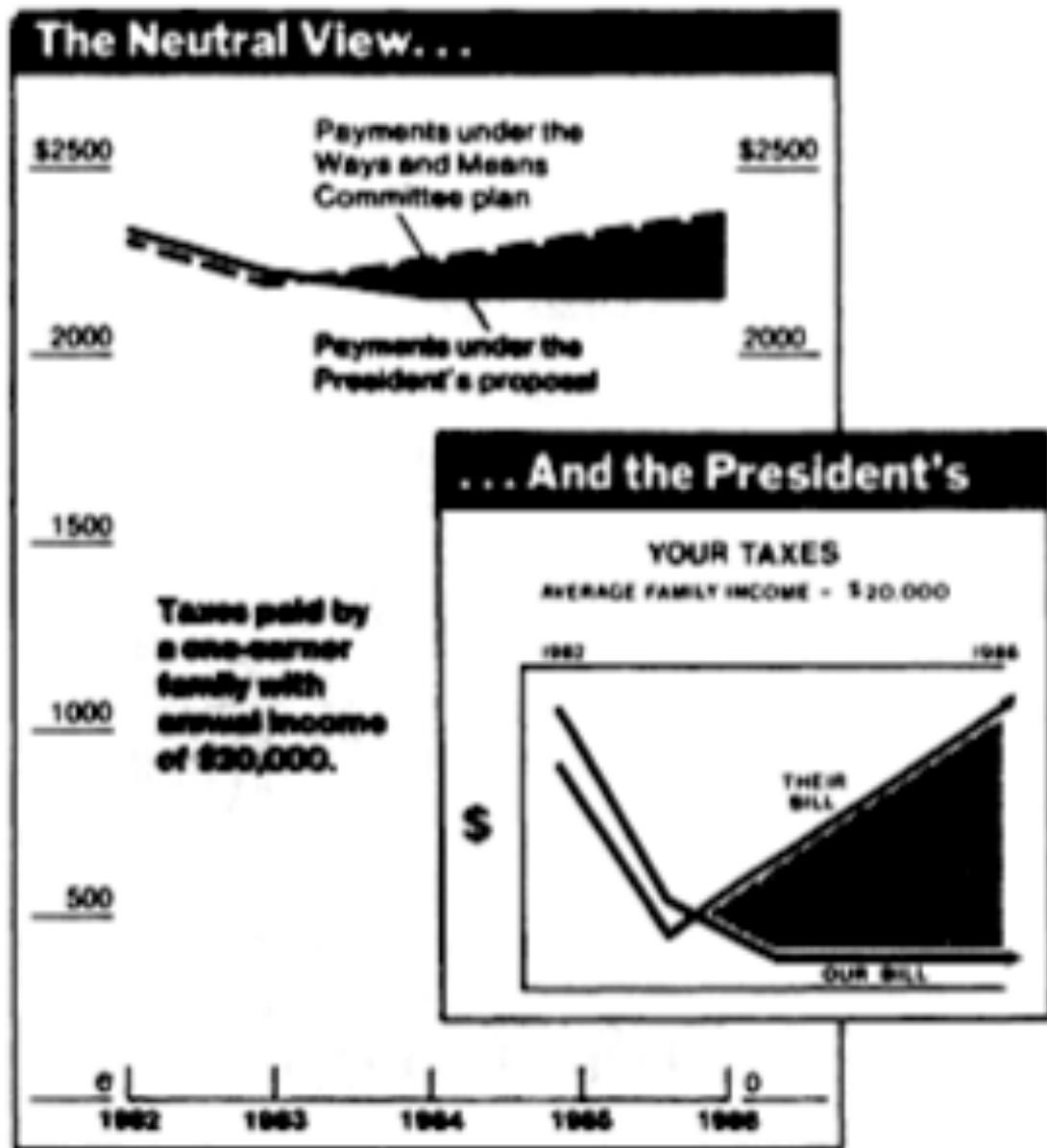
58.7%

Duty - 43.8%

VAT - 14.9%

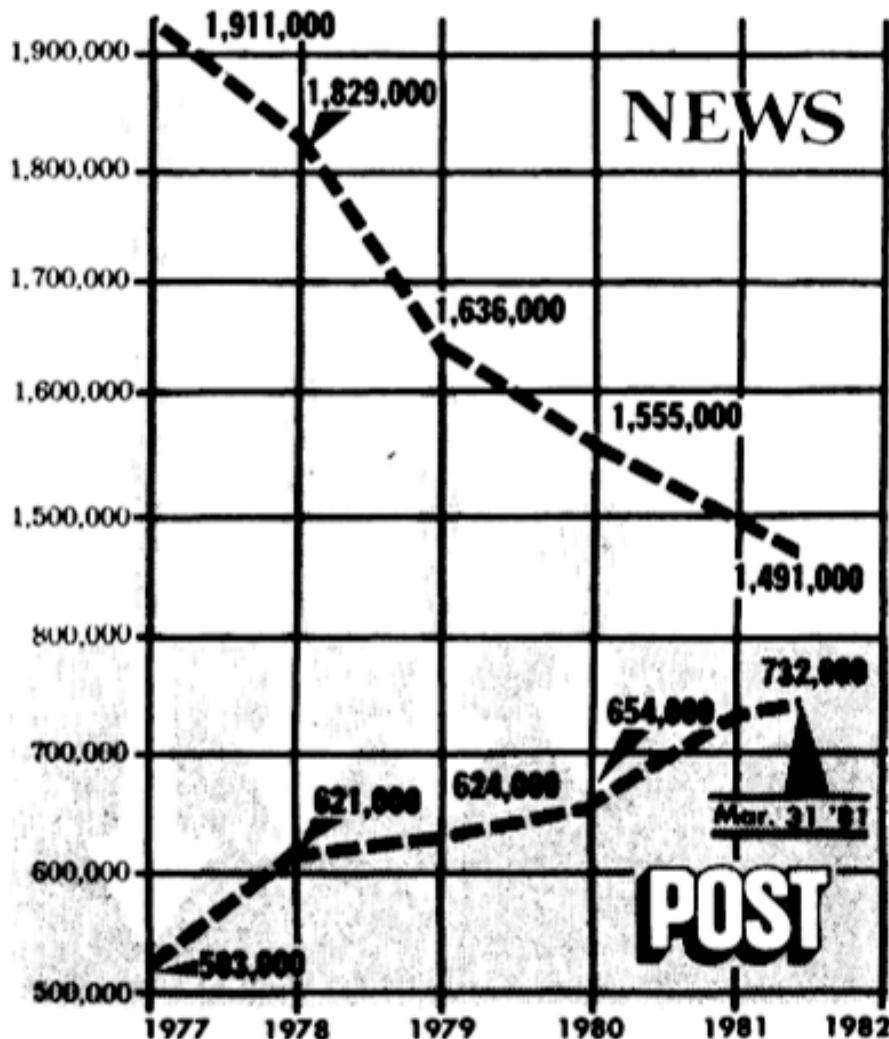
SOURCE: petrolprices.com

5. Graph data out of context.



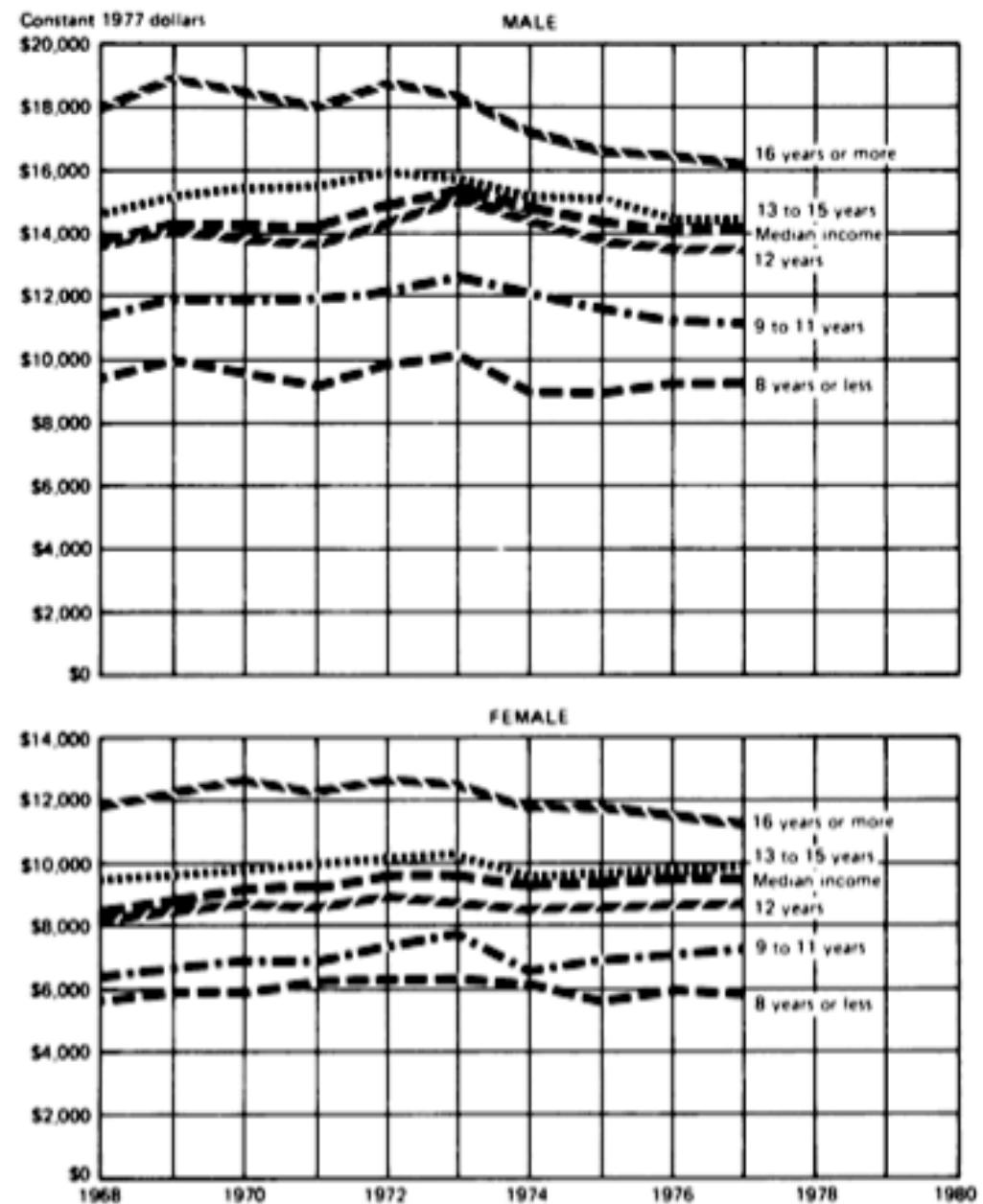
6. Change scales in mid-axis.

The soaraway Post — the daily paper New Yorkers trust

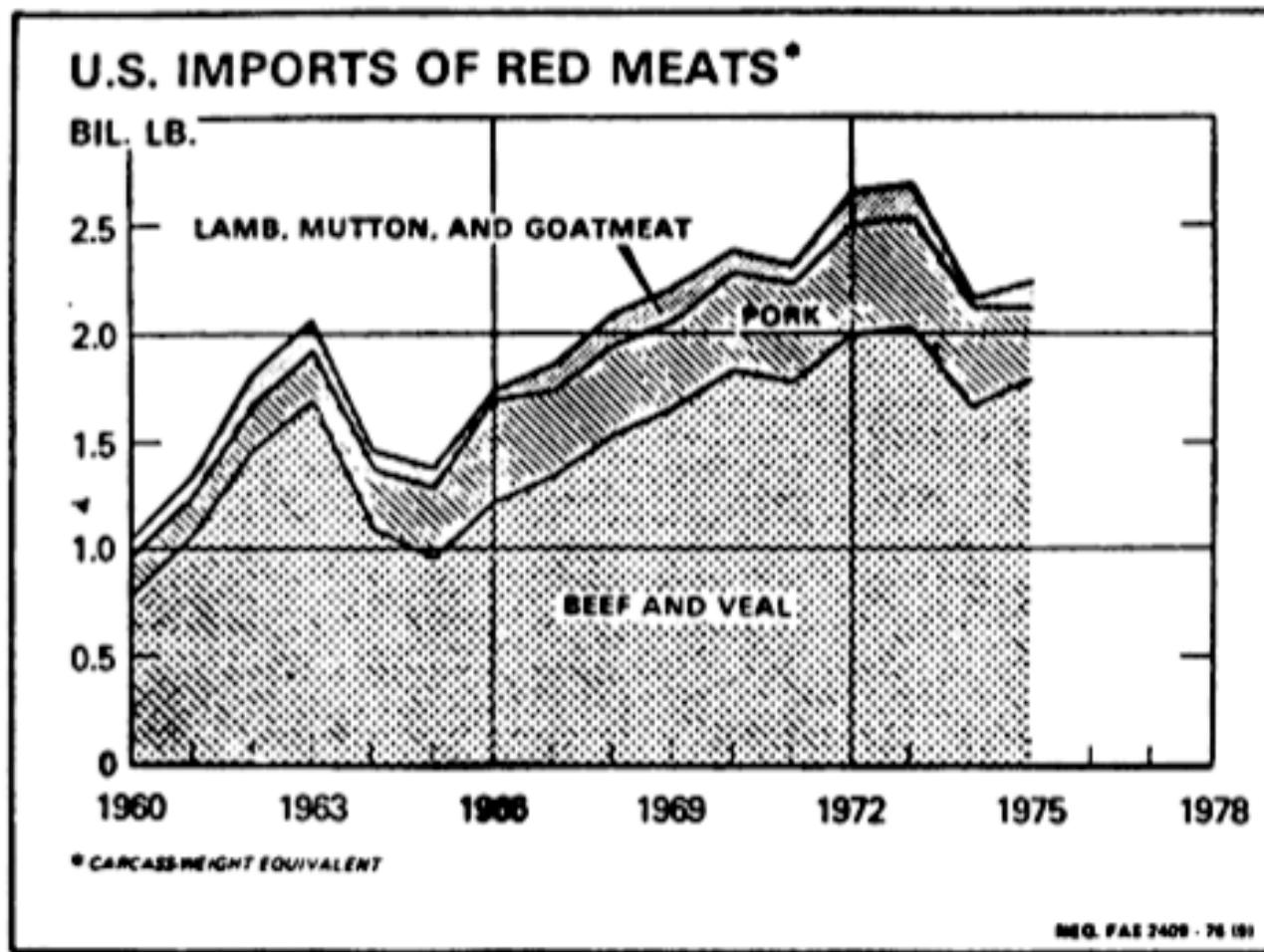


7. Emphasize the trivial. (Ignore the important.)

Median Income of Year-Round, Full-Time Workers 25 to 34 Years Old, by Sex and Educational Attainment: 1968-1977



8. Jiggle the baseline.



Sometimes varying the baseline is ok, if the main points of comparison are the first category and the total. This plot, however, is bizarre in other ways.



9. Austria first!

Insecure About Security

When it comes to Web services, how concerned are you about the following?

Confusion about the alphabet soup of standards

23% Not concerned | 63% Somewhat concerned | 14% Very concerned

Cost of new hardware

35% Not concerned | 40% Somewhat concerned | 25% Very concerned

Cost of new software

20% Not concerned | 45% Somewhat concerned | 35% Very concerned

Cost of re-engineering or integrating applications

11% Not concerned | 40% Somewhat concerned | 49% Very concerned

Cost of training

25% Not concerned | 58% Somewhat concerned | 17% Very concerned

Impact on application processing

21% Not concerned | 49% Somewhat concerned | 30% Very concerned

Impact on network bandwidth

26% Not concerned | 45% Somewhat concerned | 30% Very concerned

Security

7% Not concerned | 27% Somewhat concerned | 66% Very concerned

■ Not concerned

■ Somewhat concerned

■ Very concerned

Source: IT Architect Reader Poll

When it comes to Web services, how concerned are you about the following?

Not ■ Somewhat ■ Very Concerned

Security

7% Not concerned | 27% Somewhat concerned | 66% Very concerned

Cost of re-engineering or integrating applications

11% Not concerned | 40% Somewhat concerned | 49% Very concerned

Cost of new software

20% Not concerned | 45% Somewhat concerned | 35% Very concerned

Impact on application processing

21% Not concerned | 48% Somewhat concerned | 30% Very concerned

Confusion about the alphabet soup of standards

23% Not concerned | 63% Somewhat concerned | 14% Very concerned

Cost of training

25% Not concerned | 58% Somewhat concerned | 17% Very concerned

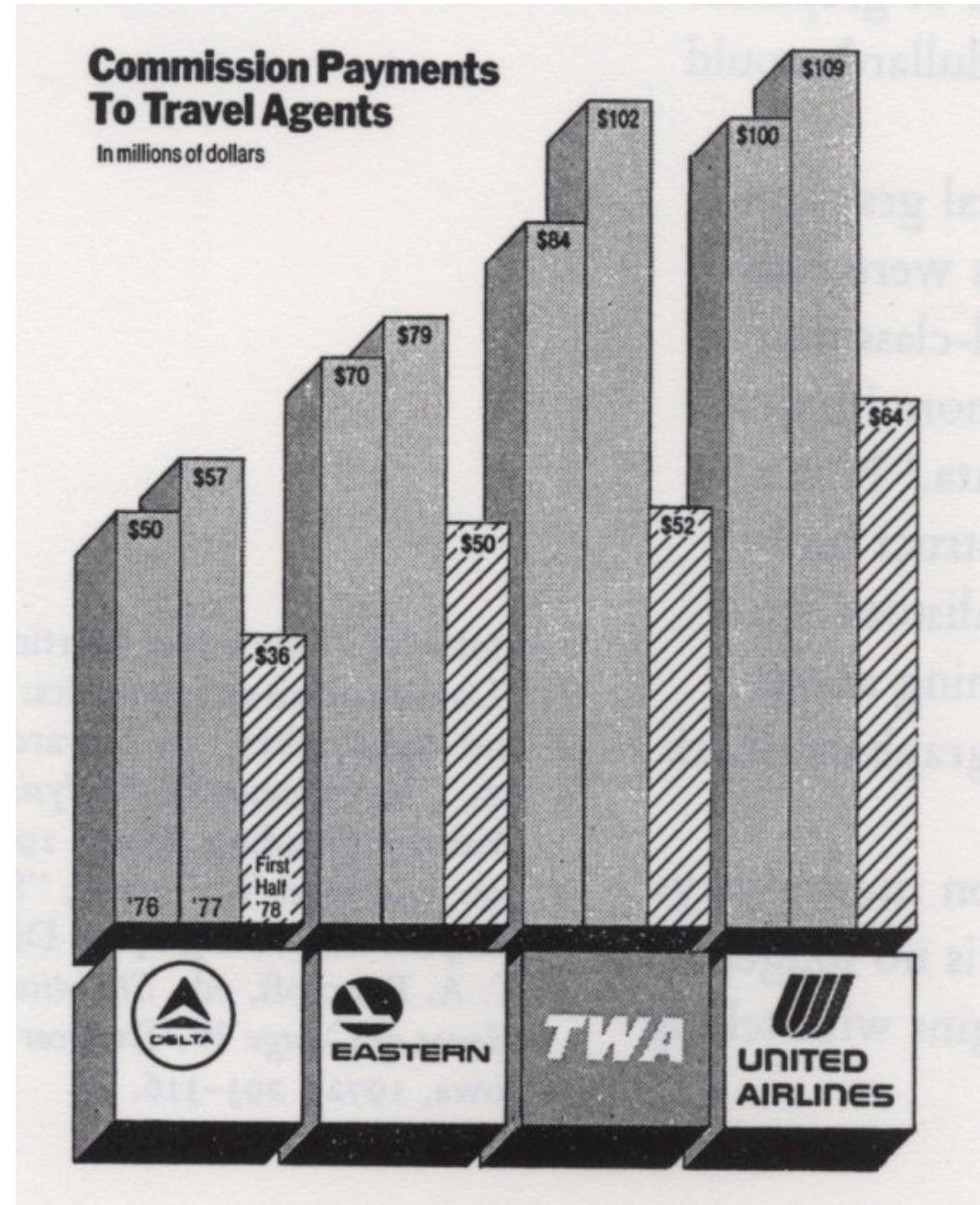
Impact on network bandwidth

26% Not concerned | 45% Somewhat concerned | 30% Very concerned

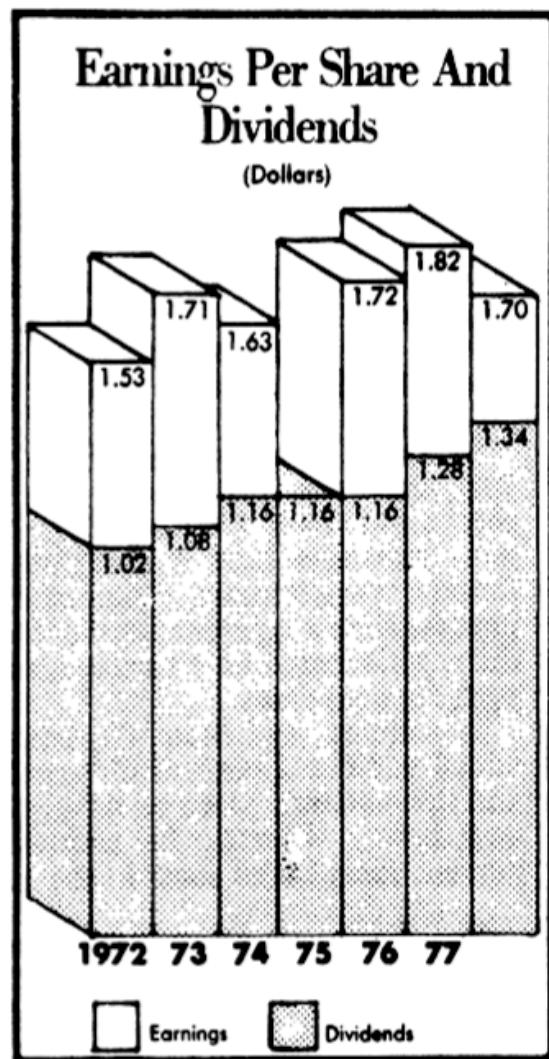
Cost of new hardware

35% Not concerned | 40% Somewhat concerned | 25% Very concerned

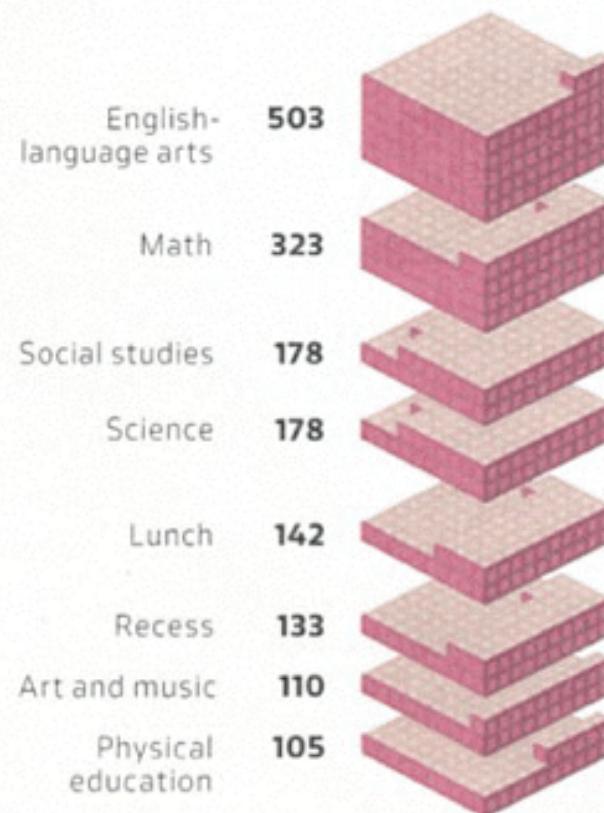
10. Label (a) Illegibly,
(b) Incompletely,
(c) Incorrectly, and
(d) Ambiguously.



III. More (dimensions) is murkier.



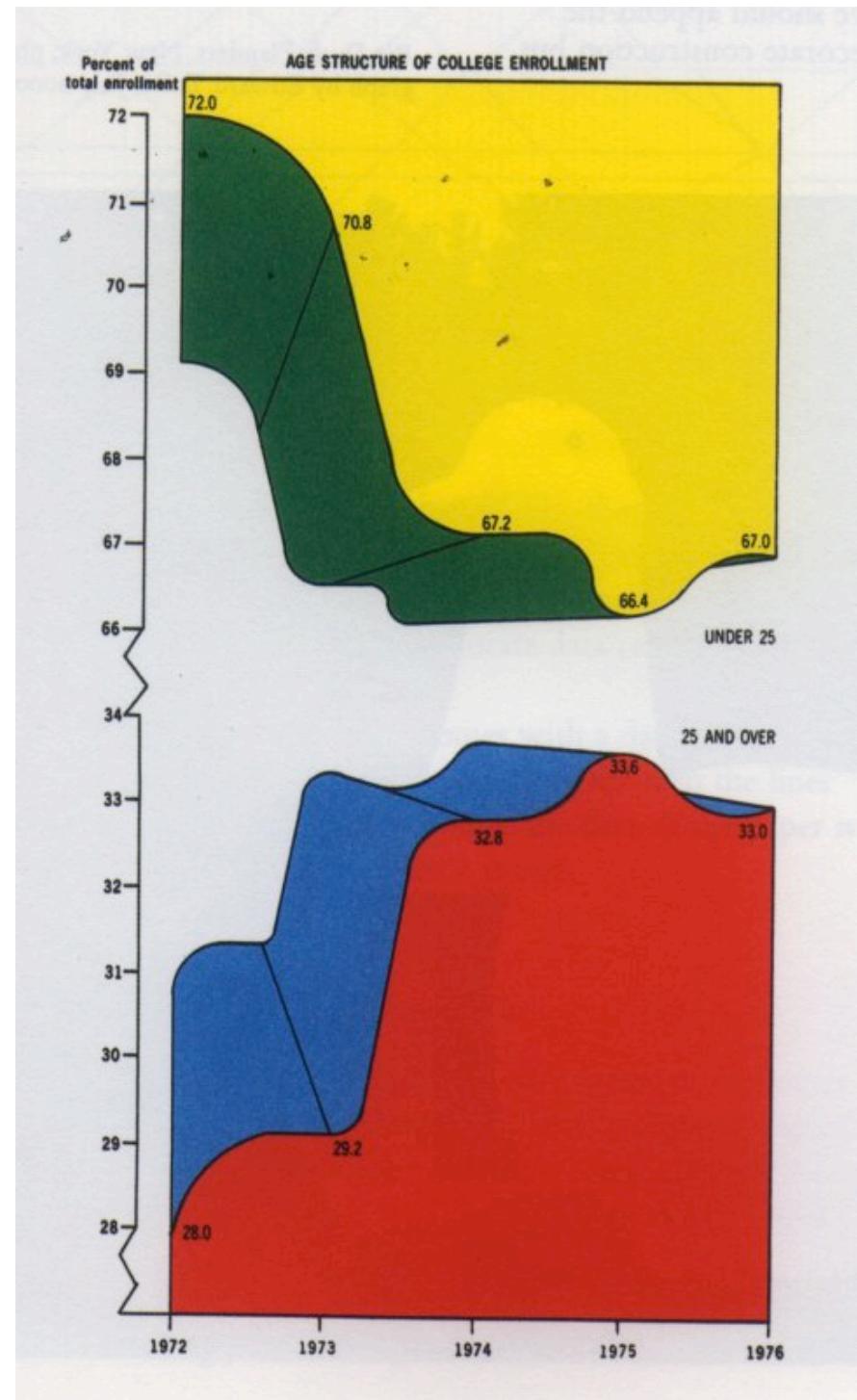
EXTRA CREDIT
Average number of minutes per week devoted to subjects or activities in elementary schools*, 2006-2007 school year



* In 349 districts surveyed nationwide

Source: Center on Education Policy
Chart by Charles M. Blow

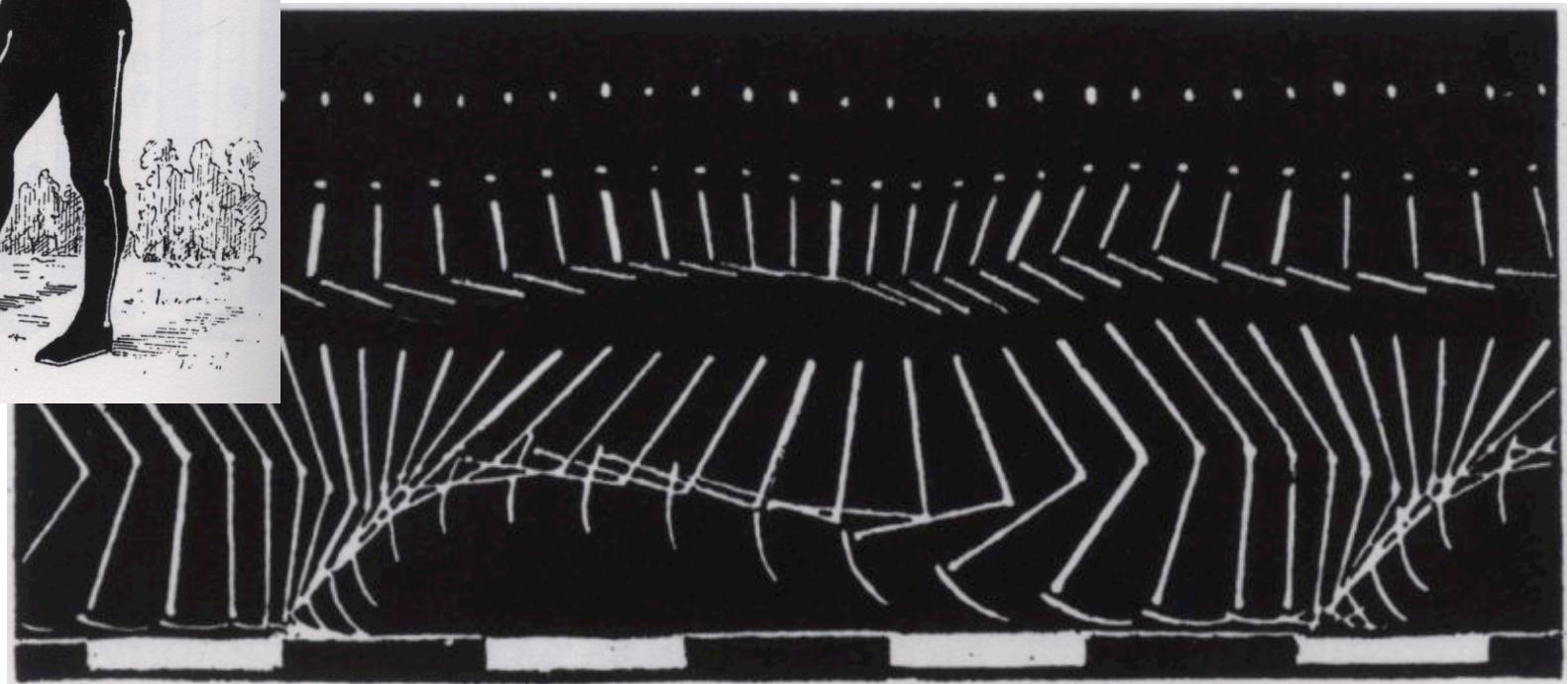
More dimensions AND colors!



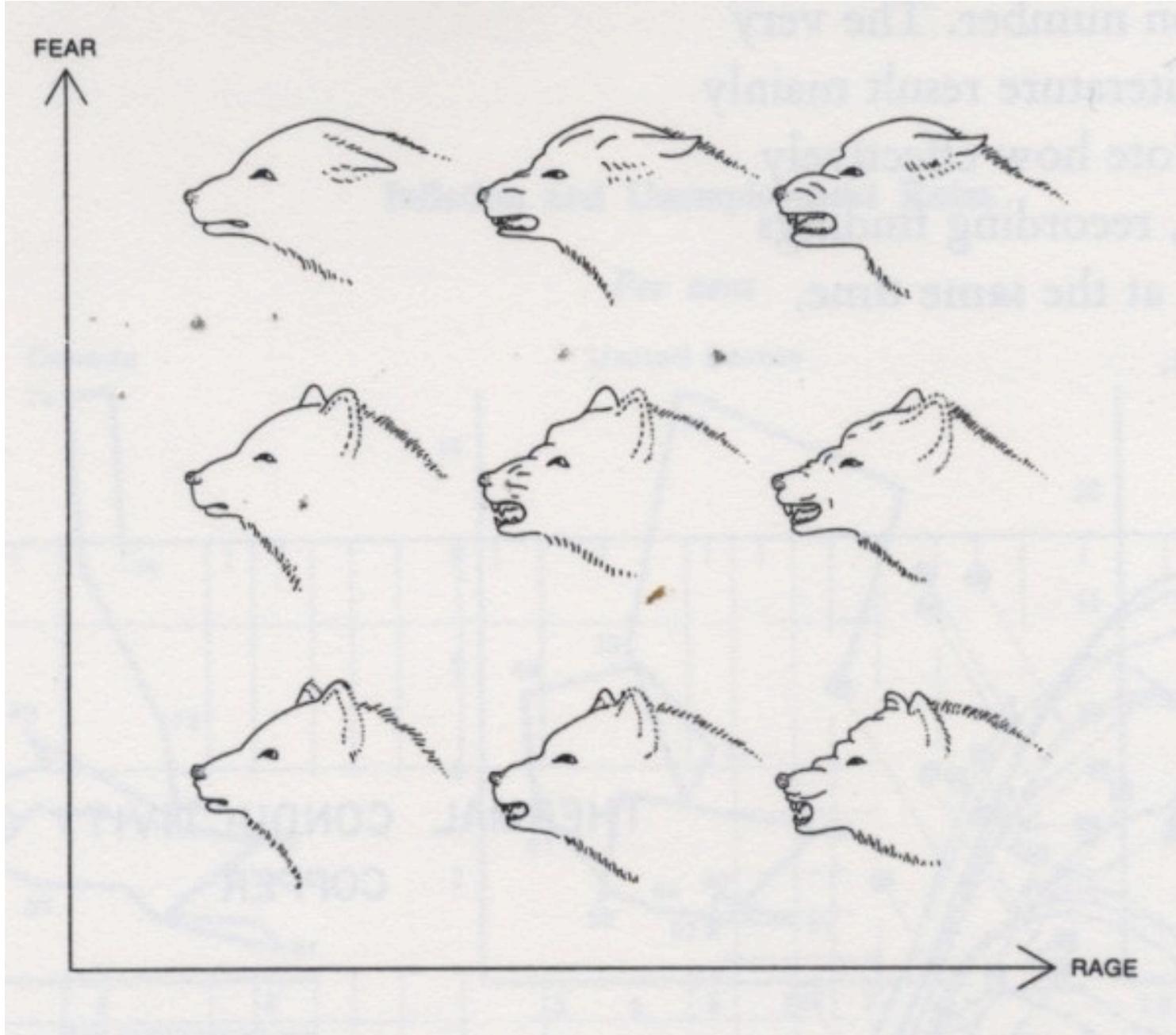
I2. If it has been done well in the past, think of another way to do it.

On the other hand, here are some creative plotting techniques you may want to consider.

I. Letting the data points represent another variable.

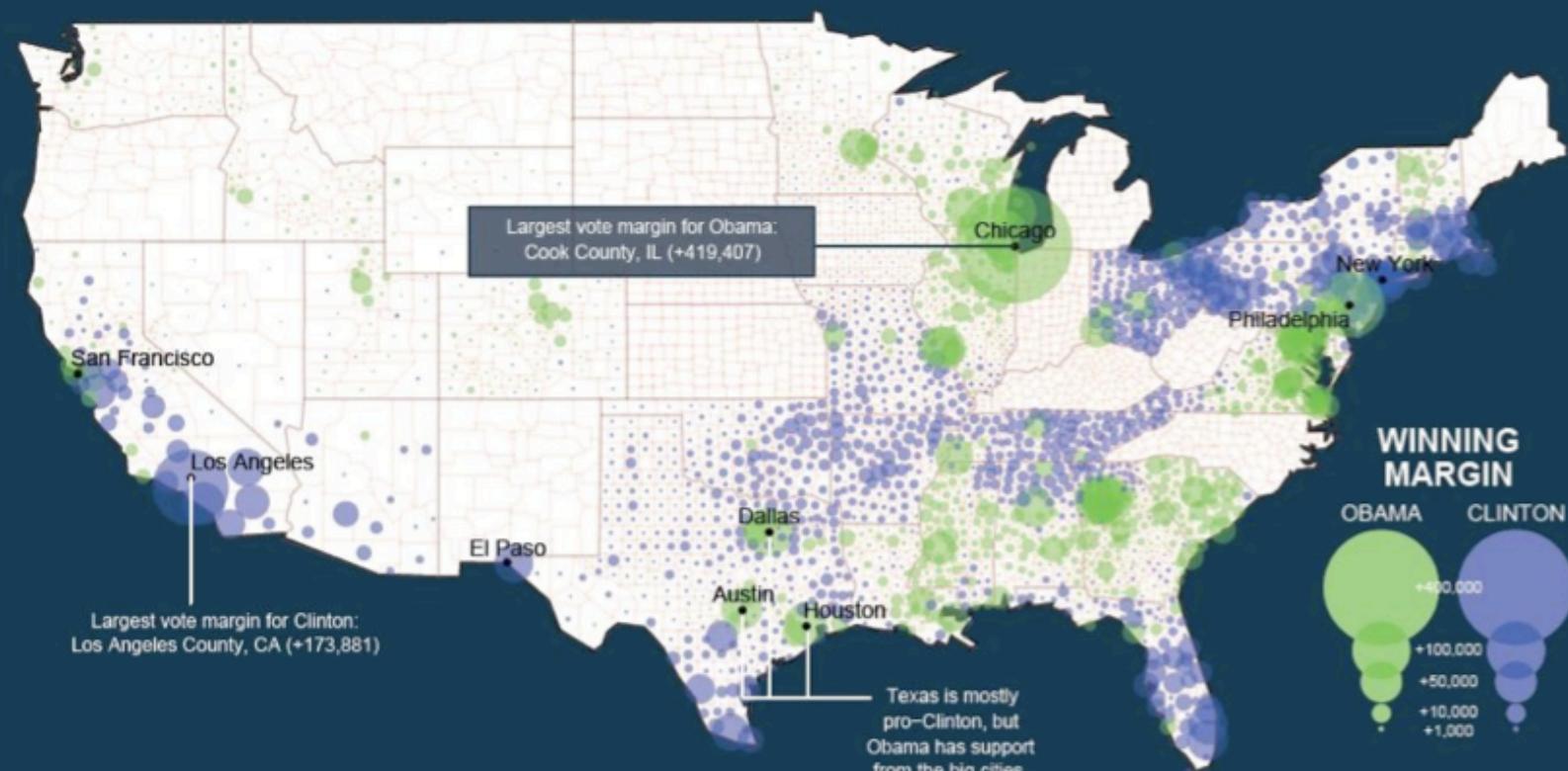


(E.J. Marey)



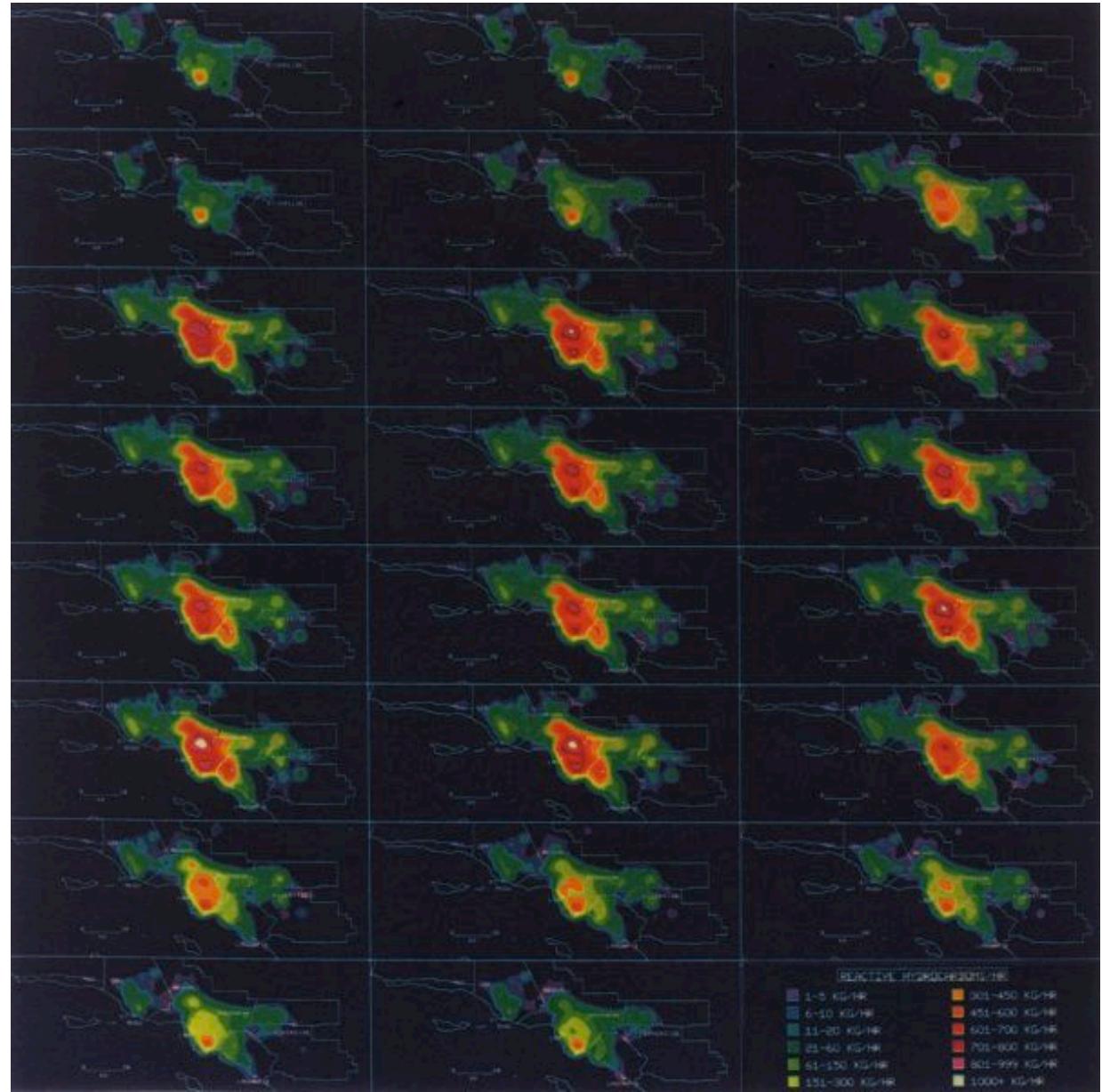
DEMOCRATIC PRIMARIES 2008

This map shows the Democratic presidential primary election results so far. The size of the circles represents how large of a winning margin one candidate had over the other. The green circles indicate a primary won by Obama, and blue circles indicate a primary won by Clinton.

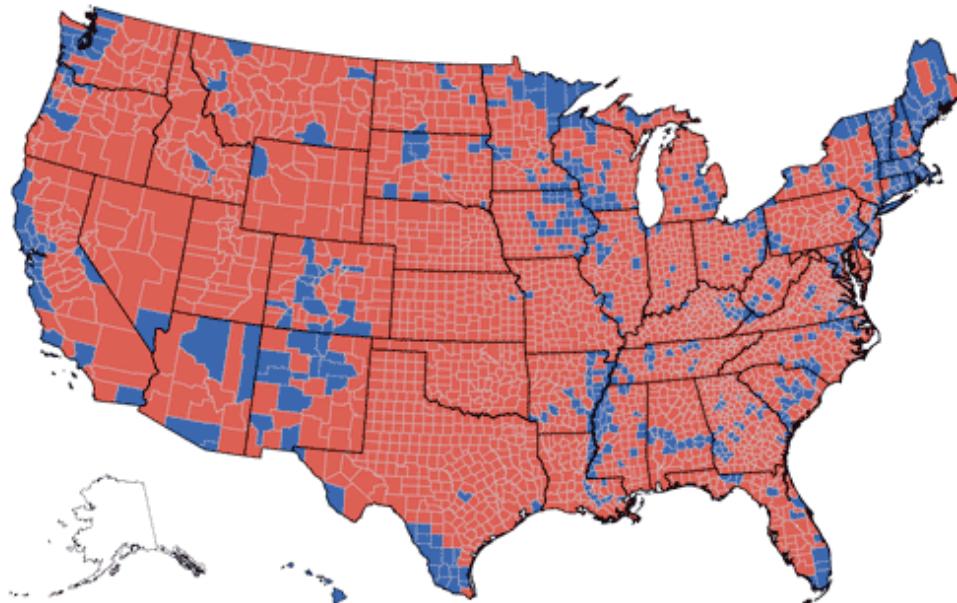


NOTE:
Primary vote data for the following states are unavailable:
Indiana, Kansas, Kentucky, Maine, Michigan, Montana, Nebraska, New Mexico,
North Carolina, North Dakota, Oregon, South Dakota, West Virginia.

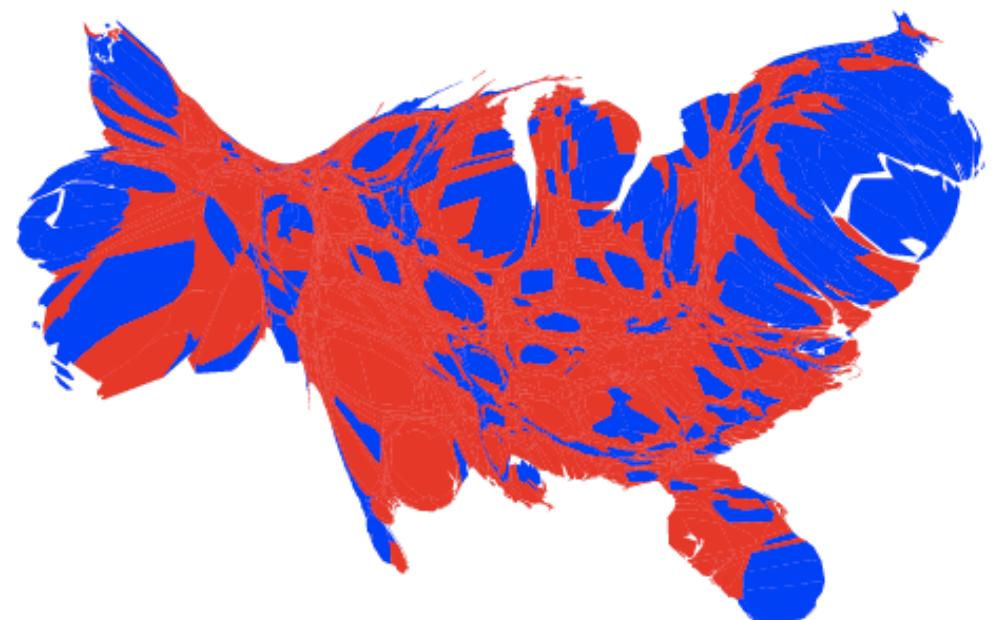
2. Using “small multiples”



3. Letting deformation represent a variable.



versus



An example from www.swivel.com

Confectionary Blog Help Feedback Sign Up! Sign In

Search

California majority party by county

By Natalie on Feb 05, 2008
Viewed 2557 times

Graph Table Map Absolute Relative All 12y 8y 2y > More Options

A bar chart comparing the number of counties in California with a majority of Democrats (light green bars) and Republicans (dark green bars) across three time points: 1995, 2000, and 2005. The y-axis represents the number of counties, ranging from 0 to 45. The x-axis shows the years 1995, 2000, and 2005. The legend indicates that light green bars represent the 'Majority of Democrats' and dark green bars represent the 'Majority of Republicans'. The data shows a shift from a majority of Democrats in 1995 to a majority of Republicans in 2005.

Year	Majority of Democrats	Majority of Republicans
1995	43	15
2000	29	29
2005	21	37

Bling Compare

Sources: California Secretary of State (<http://www.sos.ca.gov/elect...>)

Super Tuesday marks the day that 24 states and American Samoa are scheduled to hold either primary elections or caucuses in the United States. This graph shows the **number of counties in California that have a majority of Democrat or Republican registered voters in Presidential election years**. Counties with a majority of [Democratic voters](#) include: Alameda, Imperial, San Francisco, and Santa Cruz. Whereas counties with a majority of

Swivel Business
Have you tried [Swivel Business](#)?

Legend

[California Majority Party by County](#)
 [Majority of Democrats](#)
 [Majority of Republicans](#)

[More >](#)

Tags
Republican california county democrat party registered vote voters

Community Tags [Add tags](#)
no tags yet

Correlations

[Majority of Democrats](#) 100% and [Majority of Republicans](#)

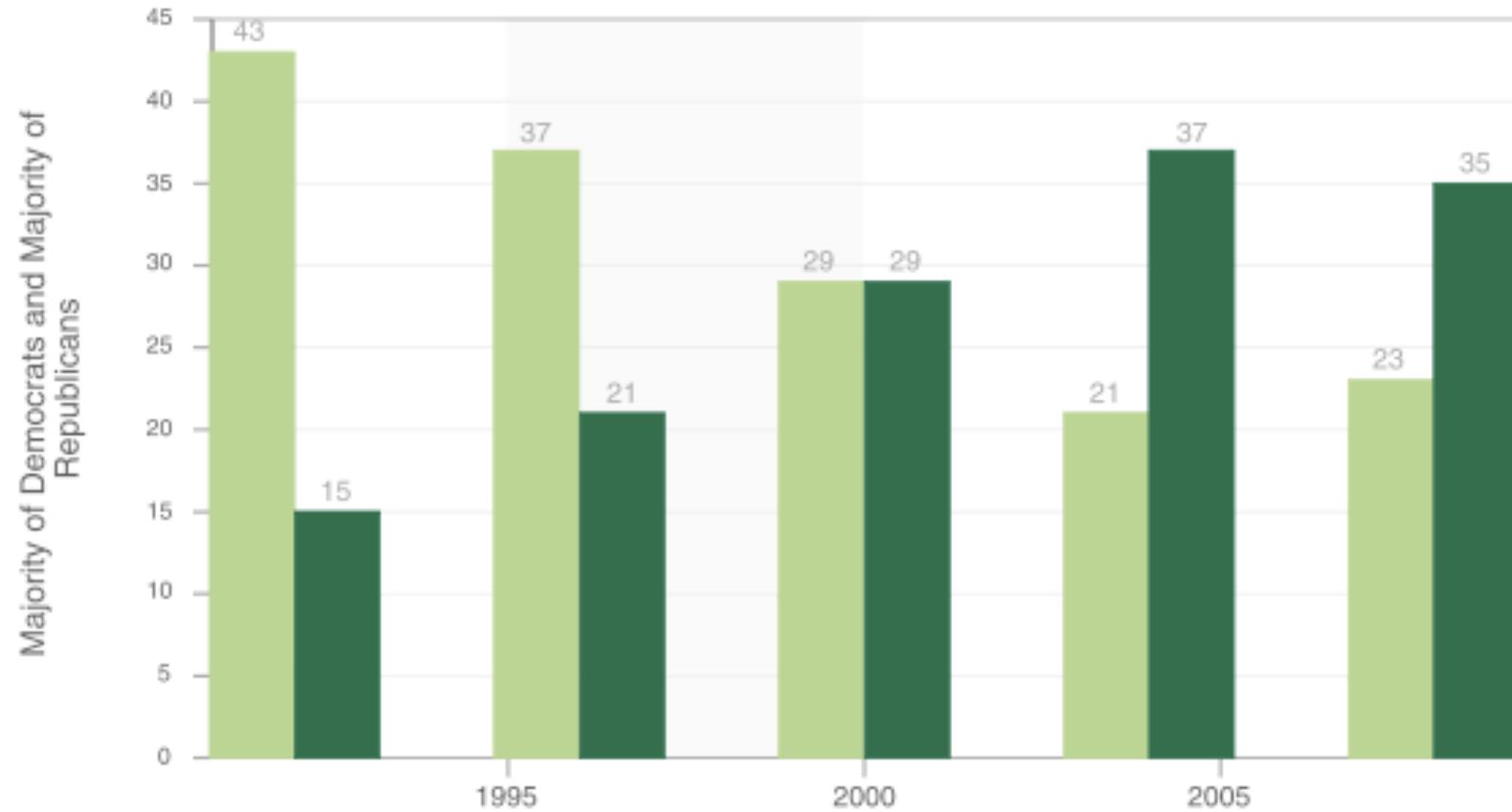
Related Graphs

[Majority of Democrats and Majority of Republicans](#)
Created By: [Natalie](#)
Views: 166

Share this Graph

Send an Email
 Post to Blog
Digg submit

Rate It



Critique:

x-axis labels poorly located - put them at election years
y-axis label misleading - these are numbers of counties
use of color could be improved (eg. red/blue)

California Counties Majority Party of Registered Voters

