Graph Construction

Statistics 133 – Fall 2014

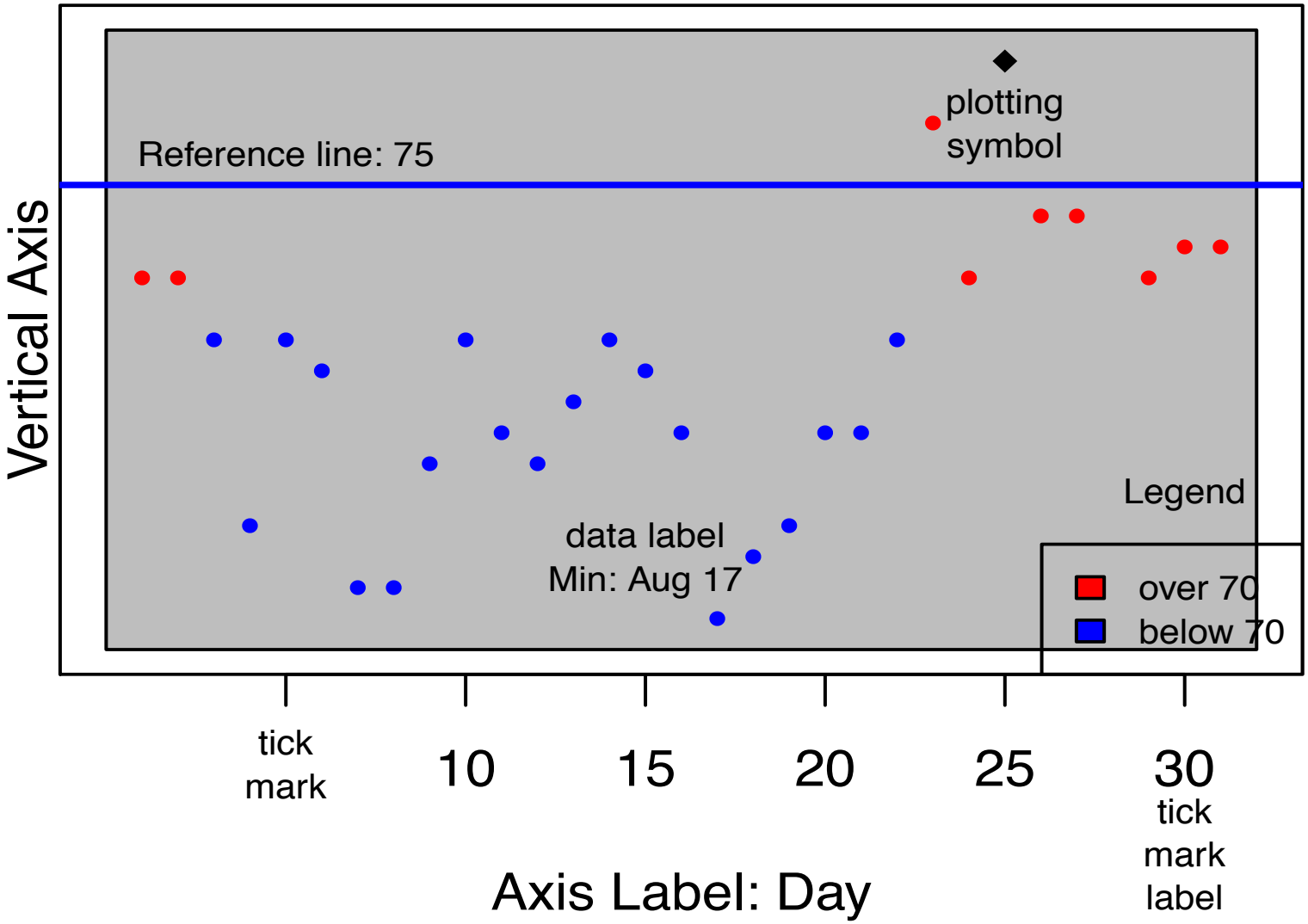
Lecture 6, Tue 9/16/14

Outline

- Vocabulary
- 3 Properties of good graph construction
  - Data stand out
  - Facilitate comparison
  - Information rich
- Perception
- Case studies

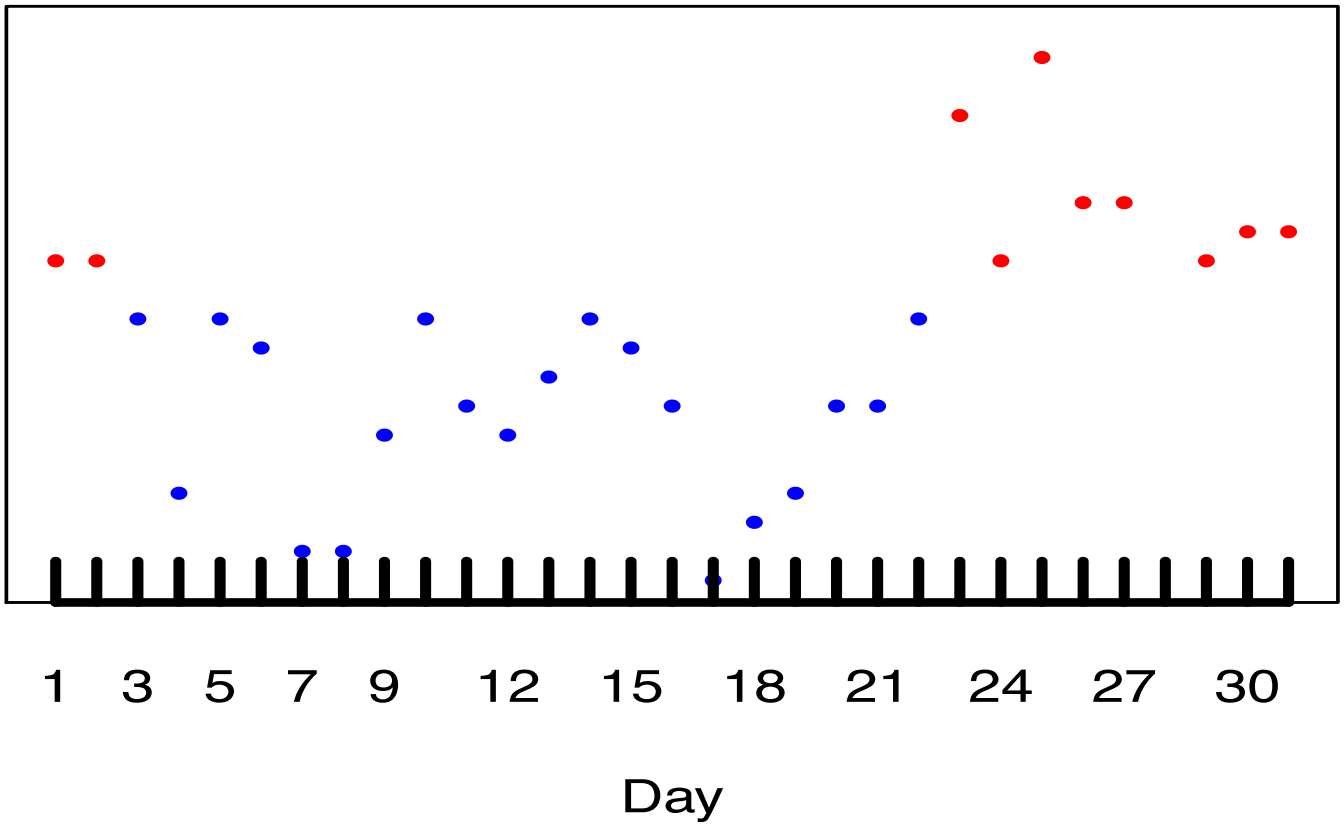
Vocabulary

Title: Temperature in August

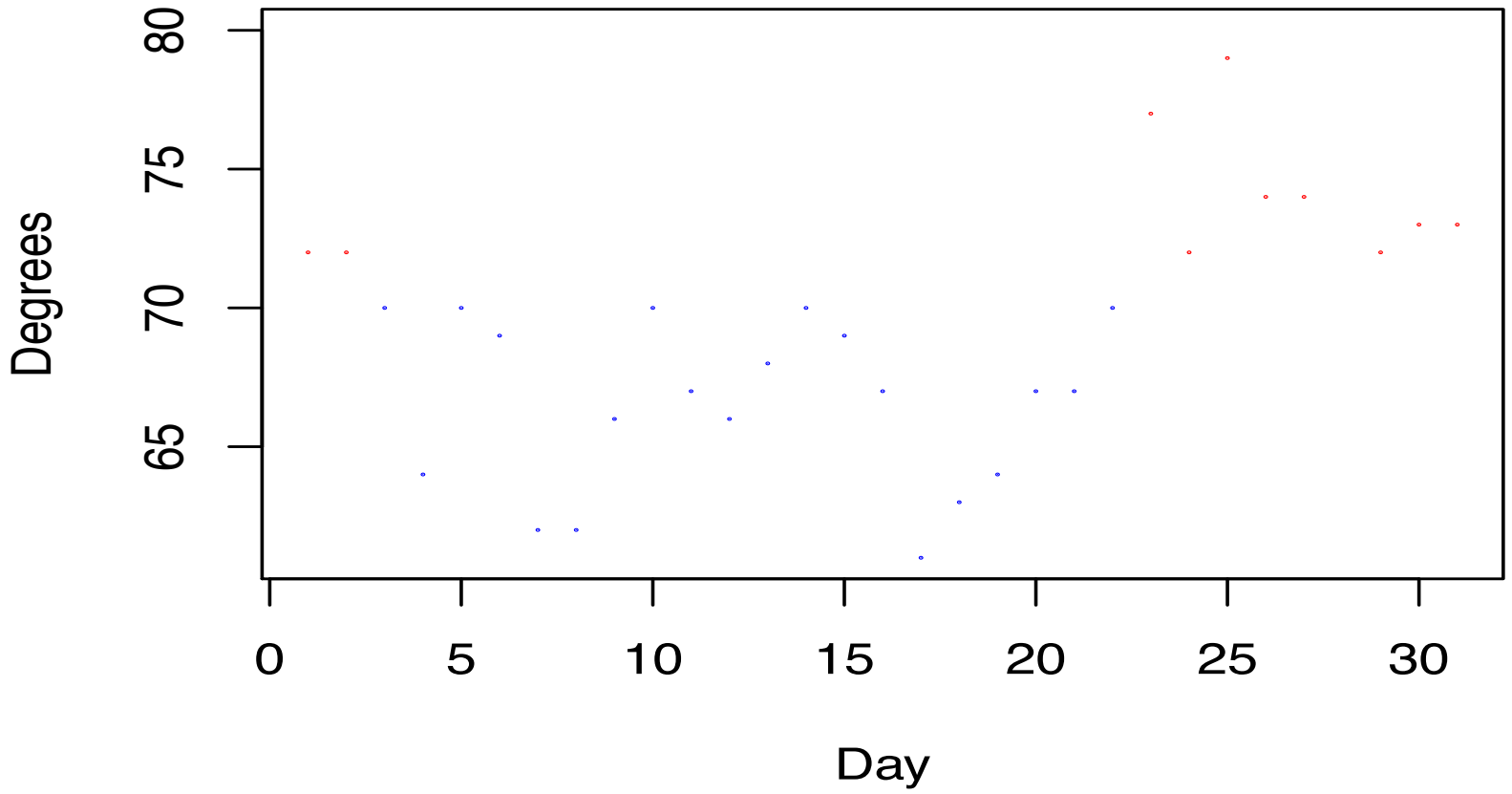


Data Stand Out

Avoid having other graph elements interfere with data

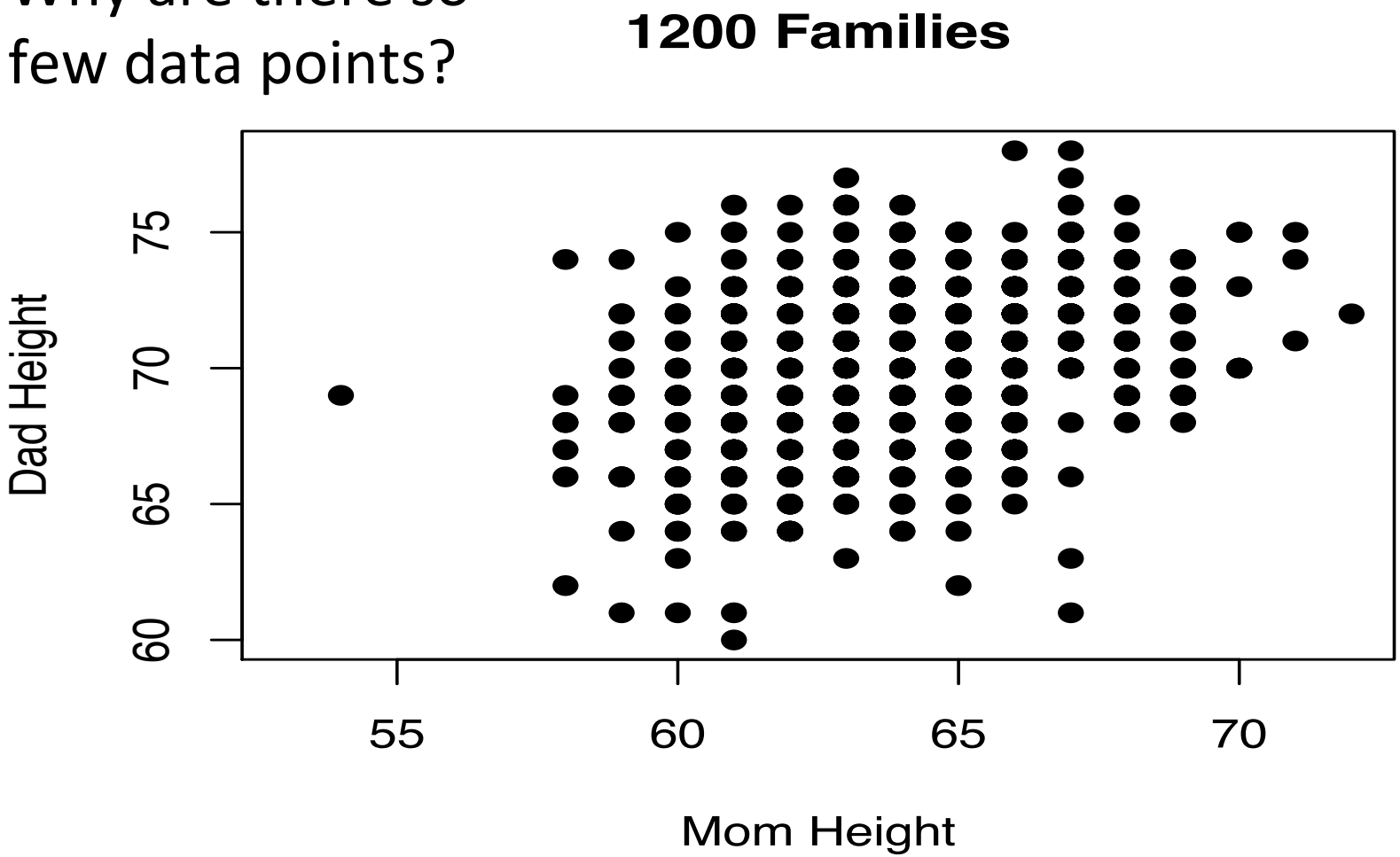


Use visually prominent symbols



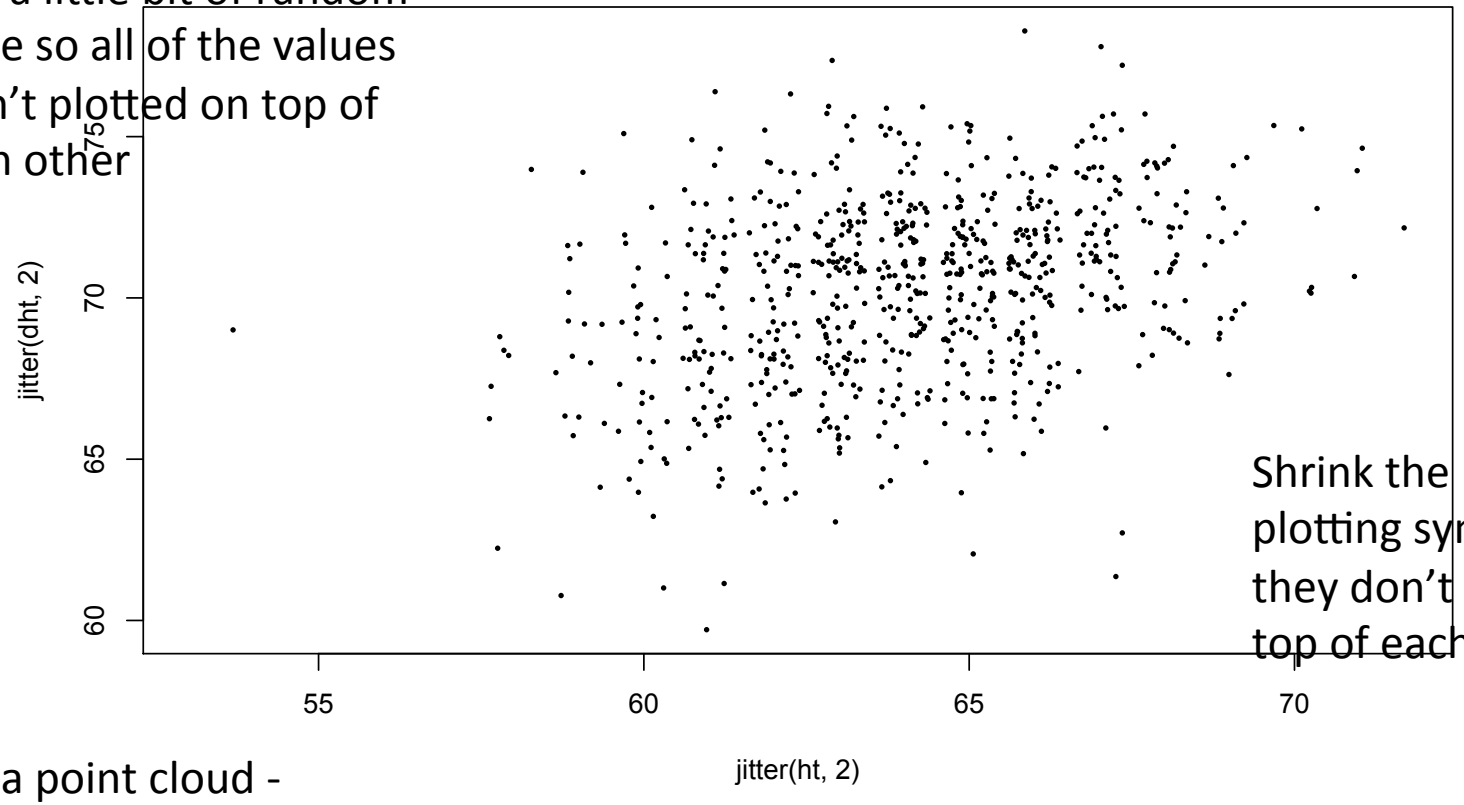
Avoid over-plotting

Why are there so few data points?



One way to avoid over plotting: Jitter the values

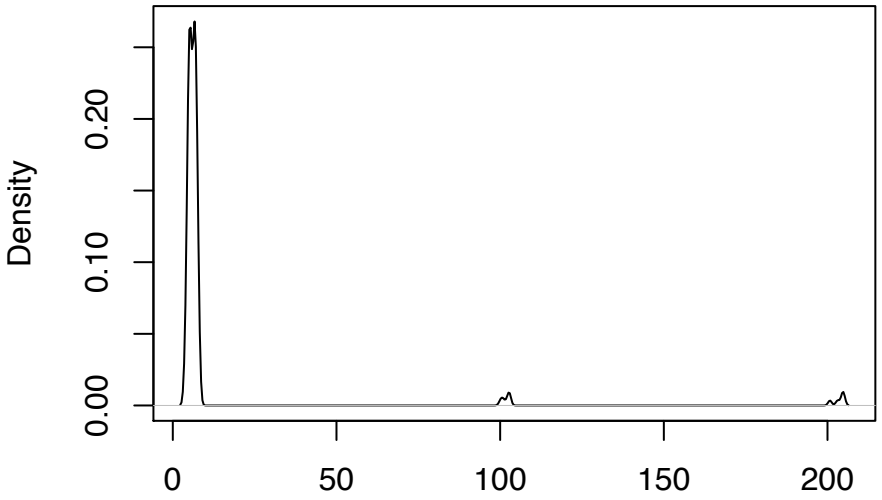
Add a little bit of random noise so all of the values aren't plotted on top of each other

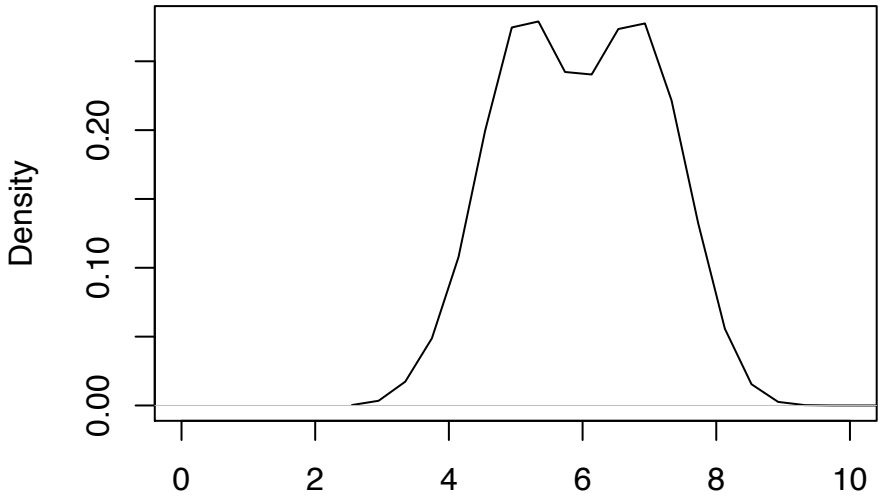


Shrink the plotting symbol so they don't plot on top of each other

See a point cloud -

Different values of data may obscure each other





Most of the data are in the 0 to 10 range.
The few large values obscure the bulk of the data.
Consider mentioning these large values in a caption, instead of showing them in the plot.

Choosing the Scale of the Axis

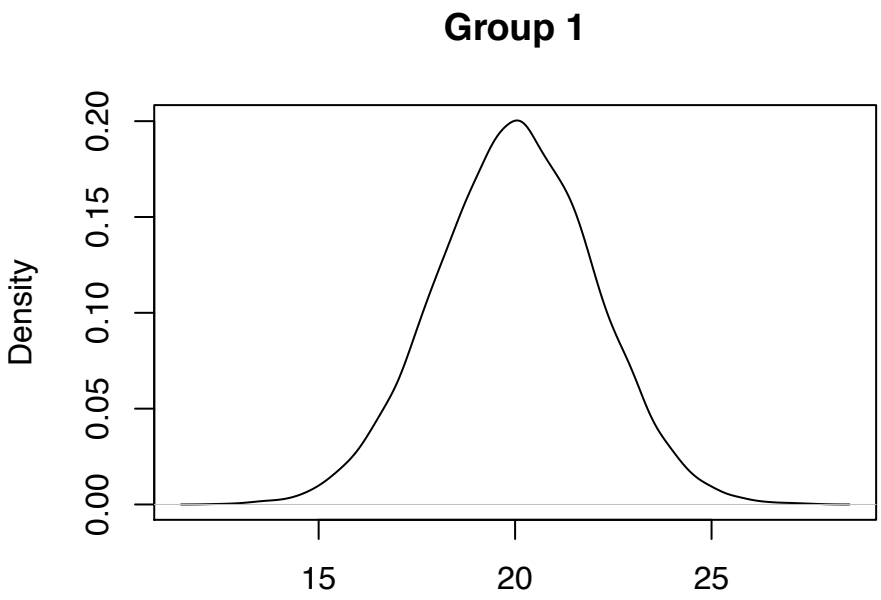
- Include all or nearly all of the data
- Fill data region
- Origin need not be on the scale
- Choose a scale that improves resolution (to be continued)

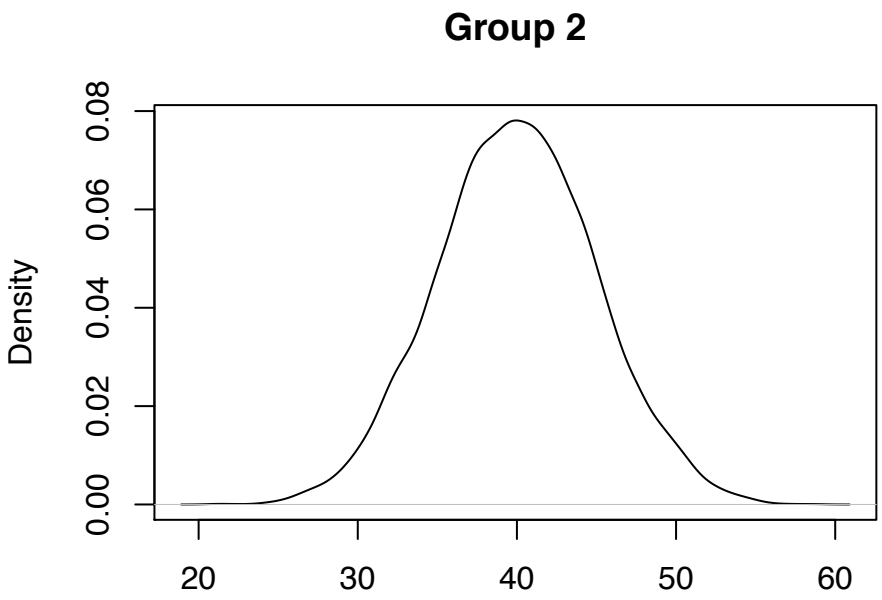
Eliminate superfluous material

- Chart junk – stuff that adds no meaning, e.g. butterflies on top of barplots, background images
- Extra tick marks and grid lines
- Unnecessary text and arrows
- Decimal places beyond the measurement error or the level of difference

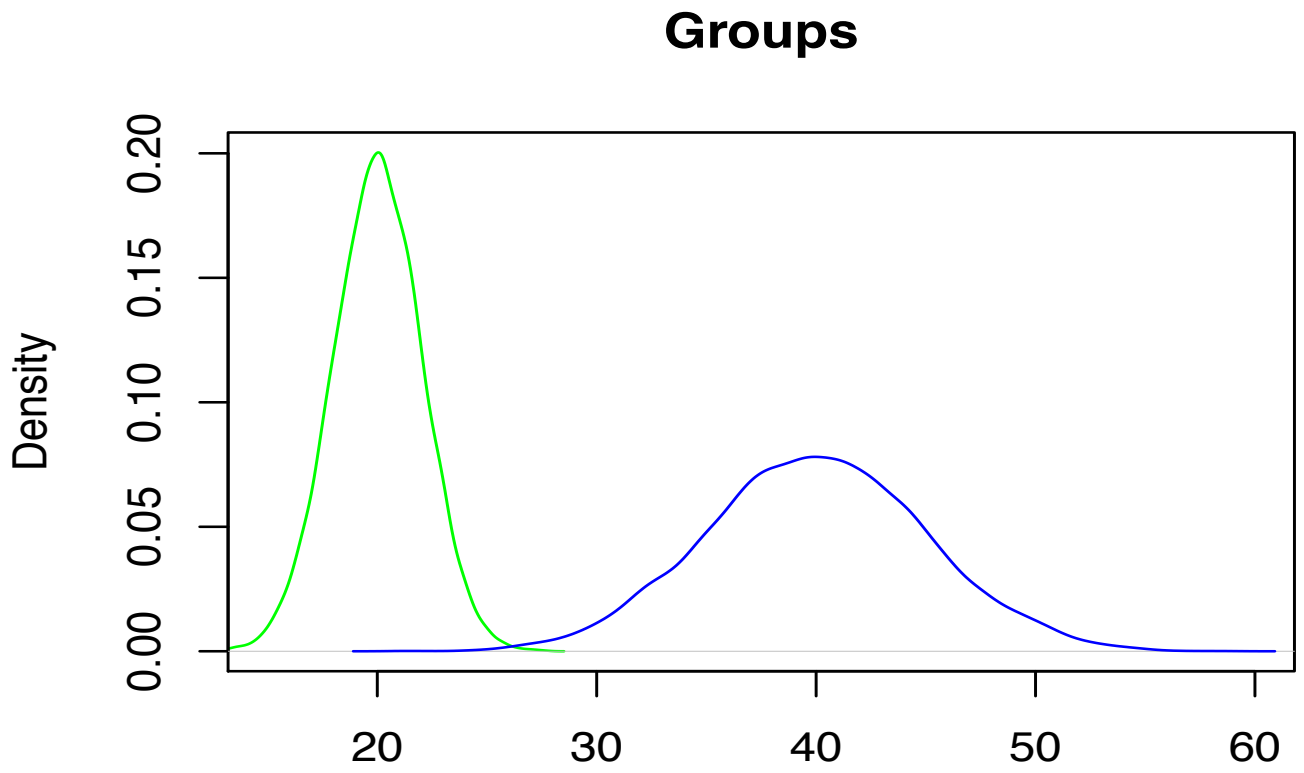
Facilitate Comparisons

Put Juxtaposed plots on same scale





Make it easy to distinguish elements of superposed plots (e.g. color)

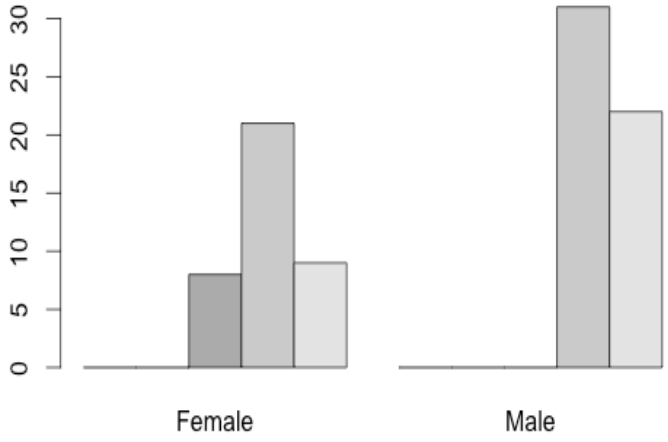


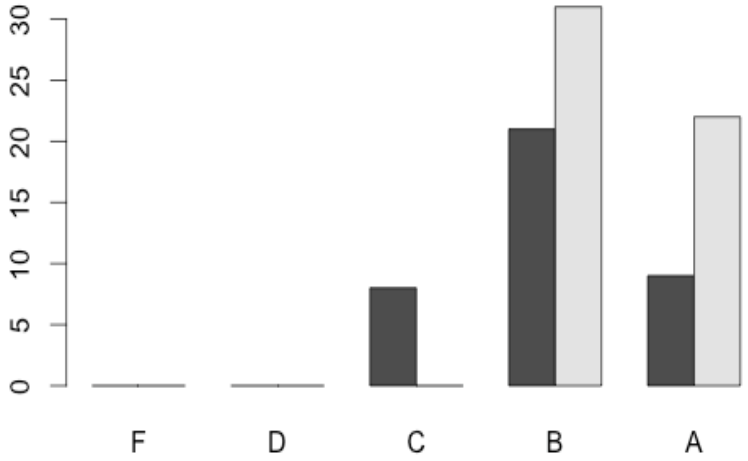
Choosing the Scale

- Keep scales on x and y axes the same for both plots to facilitate the comparison
- Zoom in to focus on the region that contains the bulk of the data
- These two principles may go counter to one another
- Keep the scale the same throughout the plot (i.e. don’t change it mid-axis)

Emphasizes the important difference

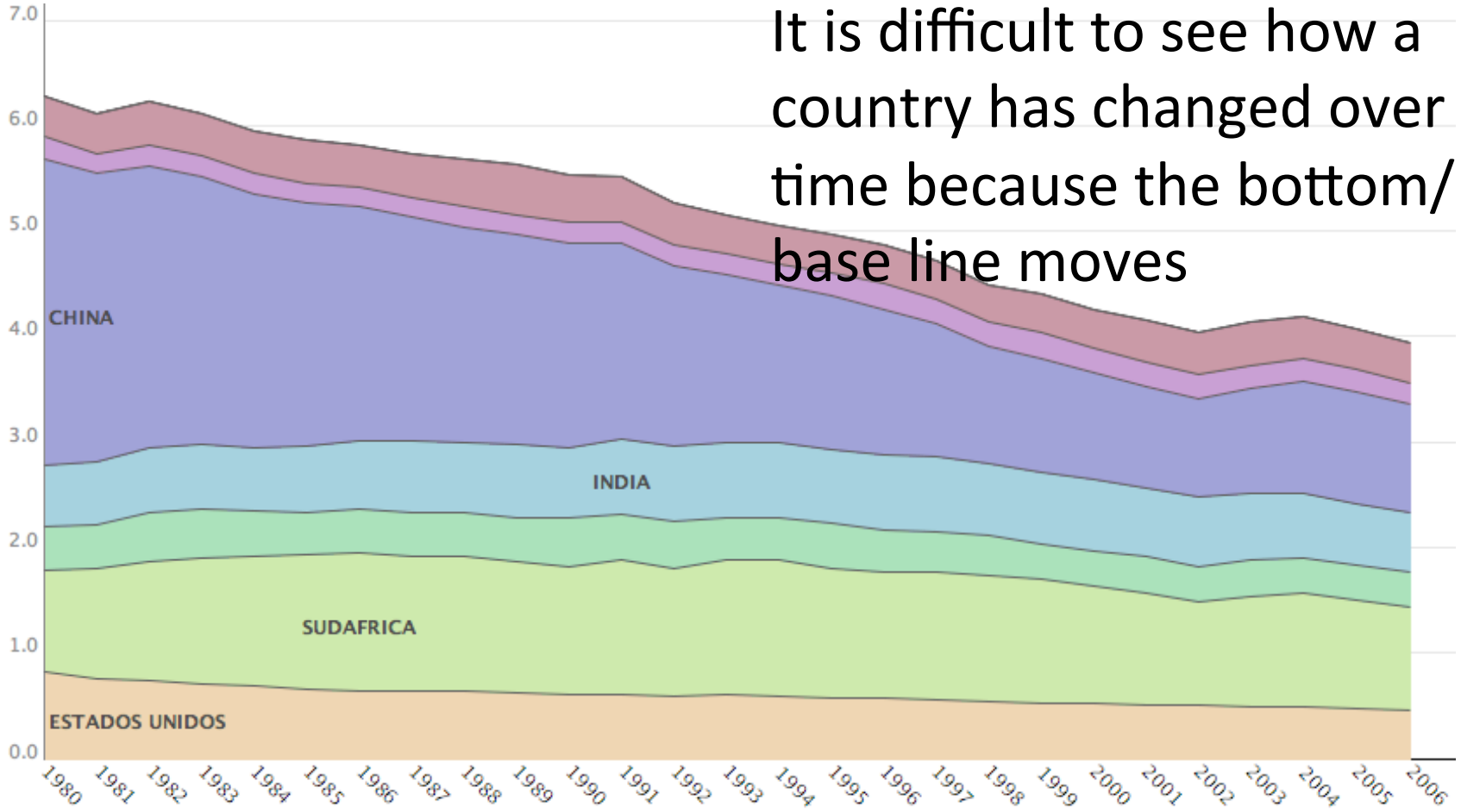
Same data on a grade survey from Stat 2 was presented in the last lecture.





Which of these side-by-side bar plots emphasizes the important?

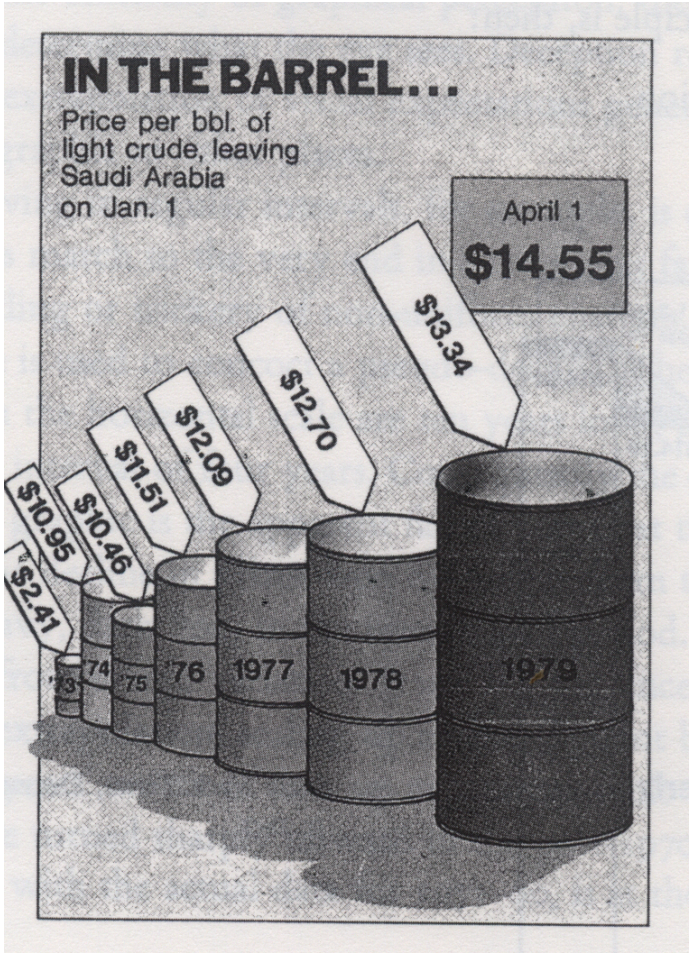
Avoid Jiggling the baseline



It is difficult to see how a country has changed over time because the bottom/ base line moves

Comparison: volume, area, height

We naturally compare the volume of the barrels, but the change is really the height of the barrels



Information Rich

How to make a plot information rich

- Describe what you see in the **Caption**
- Add context with **Reference Markers** (lines and points) including text
- Add **Legends** and **Labels**
- Use color and plotting symbols to add more information
- Plot the same thing more than once in different ways/scales
- Reduce clutter

Captions

- Captions should be comprehensive
- Captions should be self-contained
- Captions should:
  - Describe what has been graphed
  - Draw attention to important features
  - Describe conclusions drawn from graph

Good Plot Making Practice

- Put major conclusions in graphical form
- Provide reference information
- Proof read for clarity and consistency
- Graphing is an iterative process
- Multiplicity is OK, i.e. two plots of the same variable may provide different messages
- Make plots data rich

Perception

Color, shape (including banking) can affect your ability to make good comparisons

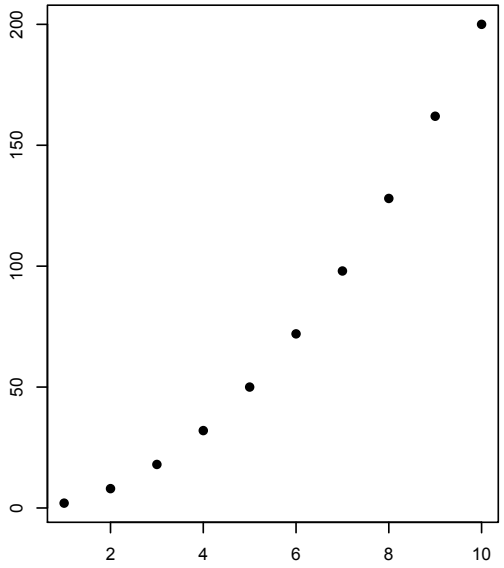
Banking: Aspect Ratio

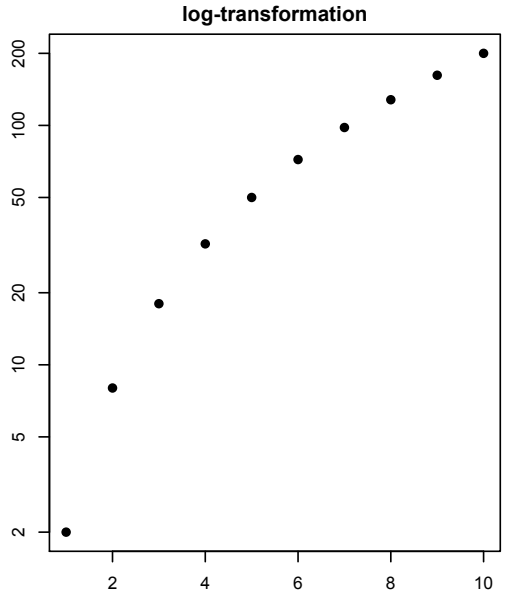
- The height/width of the data region was selected to be about 1 so that the trend line is at about 45 degrees.
- The Aspect ratio affects our visual decoding of the rate of change
- The banking to 45 degrees helps us see rate of change
- The ability to effectively judge rate of change allows us to see important patterns in data

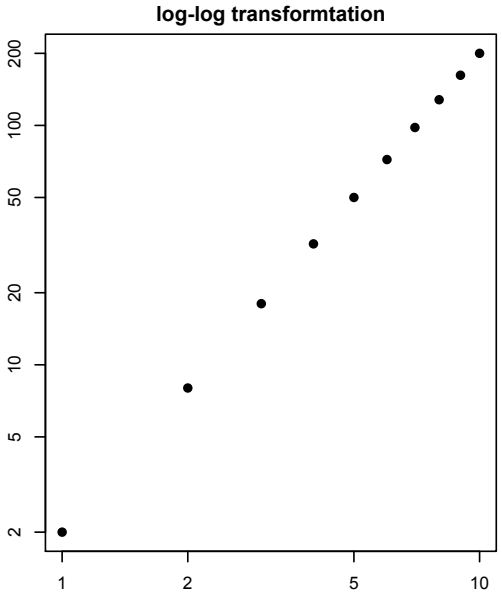
Banking at 45 degrees

- Roughly: Examine the absolute value of the orientation of segments, they should be centered at 45 degrees.
- Transformations to improve the aspect ratio uncovers the structure of the relationship between variables
- Easier to see important features

Bank to 45 degrees







Shapes

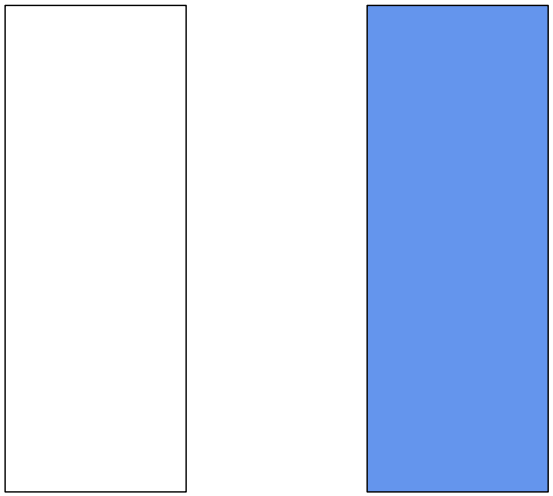
POP QUIZ!!!

Open a text file
or
Number your paper 1-6

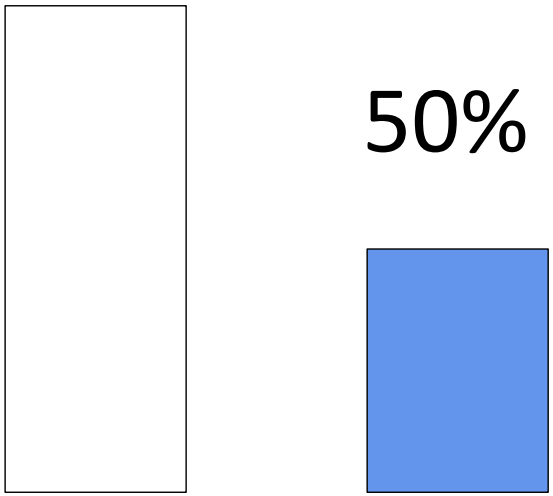
1. _____
2. _____
3. _____
4. _____
5. _____
6. _____

Warm up:
What % of the white is the blue?

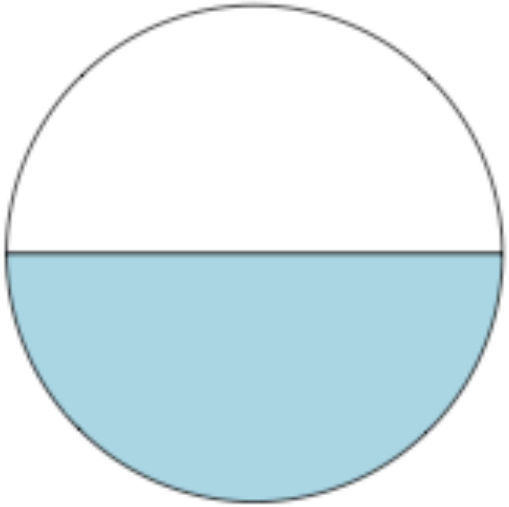
100%



Warm up:
What % of the white is the blue?

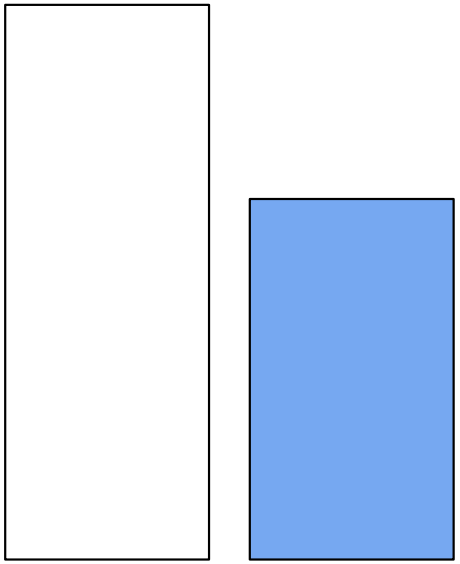


Warm up:
What percent of the white is the blue?

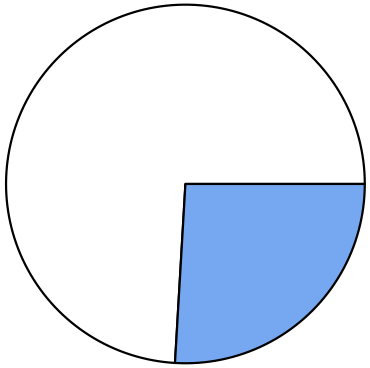


100%

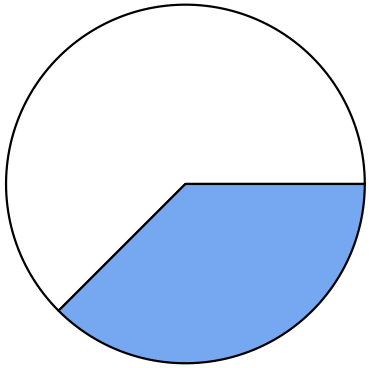
1. What % of the white is the blue?



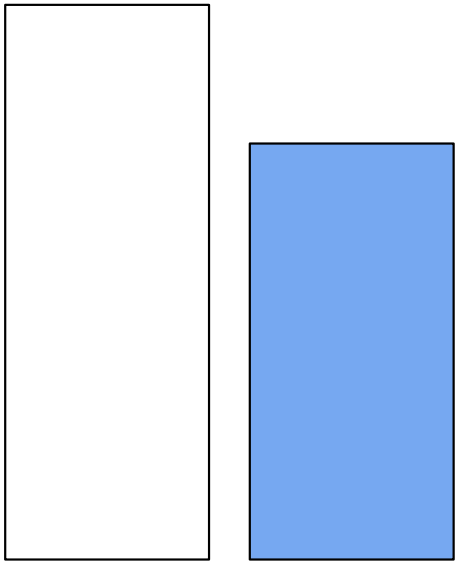
2 What % of the white is the blue?



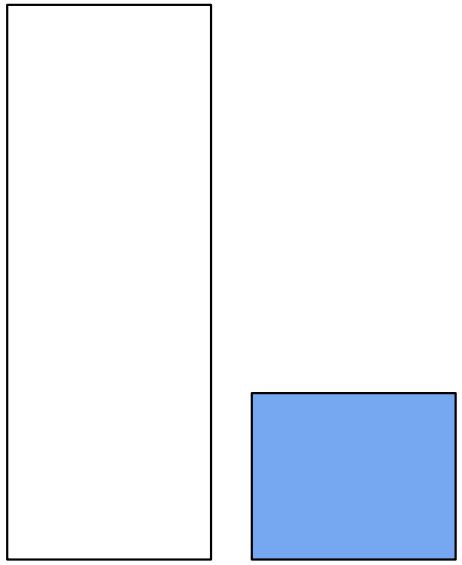
3. What % of the white is the blue?



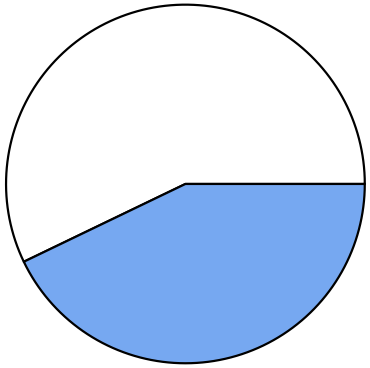
4. What % of the white is the blue?



5. What % of the white is the blue?



6 What % of the white is the blue?



How accurate were you?

You Guess	Truth	Absolute Error	Type
1. __70__	65	5	Bar
2. __33__	35	2	Pie
3. __75__	60	15	Pie
4. __75__	75	0	Bar
5. __35__	30	5	Bar
6. __85__	75	10	Pie

Bar plot vs Pie chart

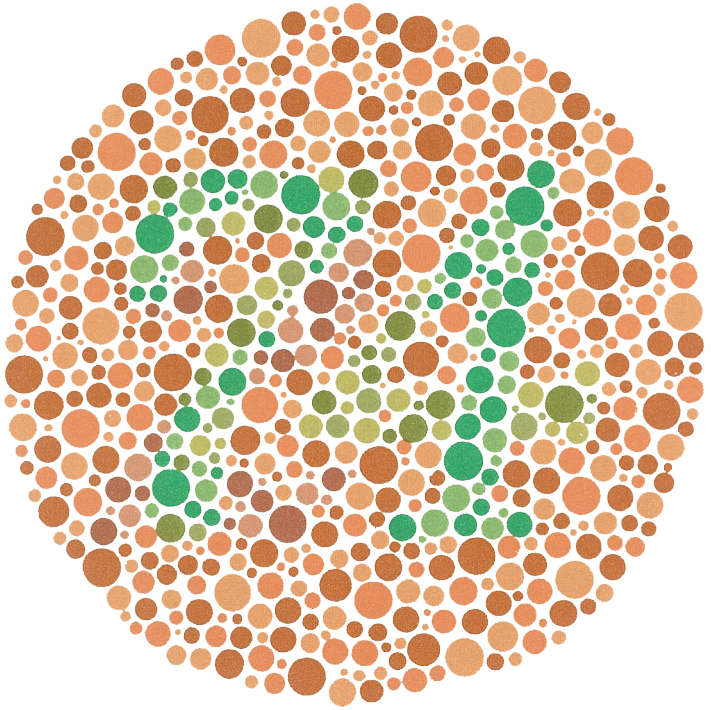
- Cleveland's experiment had a group of subjects judge 40 pairs of values on bar chars and the same 40 pairs on pie charts: **What percent the smaller was of the larger?**
- Pie chart judgments are less accurate than bar chart judgments
- Bar chart errors are about the same size for all percents.
- Pie chart errors tend to be larger for percents greater than 35%

Color

Color Guidelines

- Choosing a set of colors which work well together is a challenging task for anyone who does not have an intuitive gift for color

- 7-10% of males are red-green color blind.

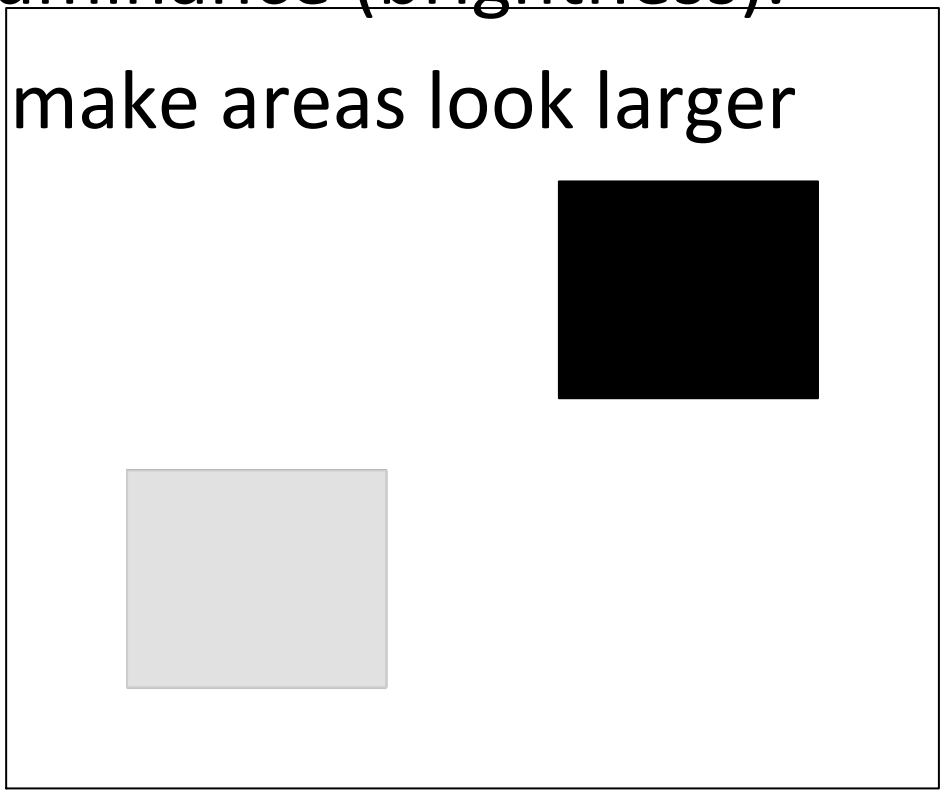


Colorfulness

- Saturated/colorful colors are hard to look at for a long time.
- They tend to produce an after-image effect which can be distracting.

Luminance

- If the size of the areas presented in a graph is important, then the areas should be rendered with colors of similar luminance (brightness).
- Lighter colors tend to make areas look larger than darker colors



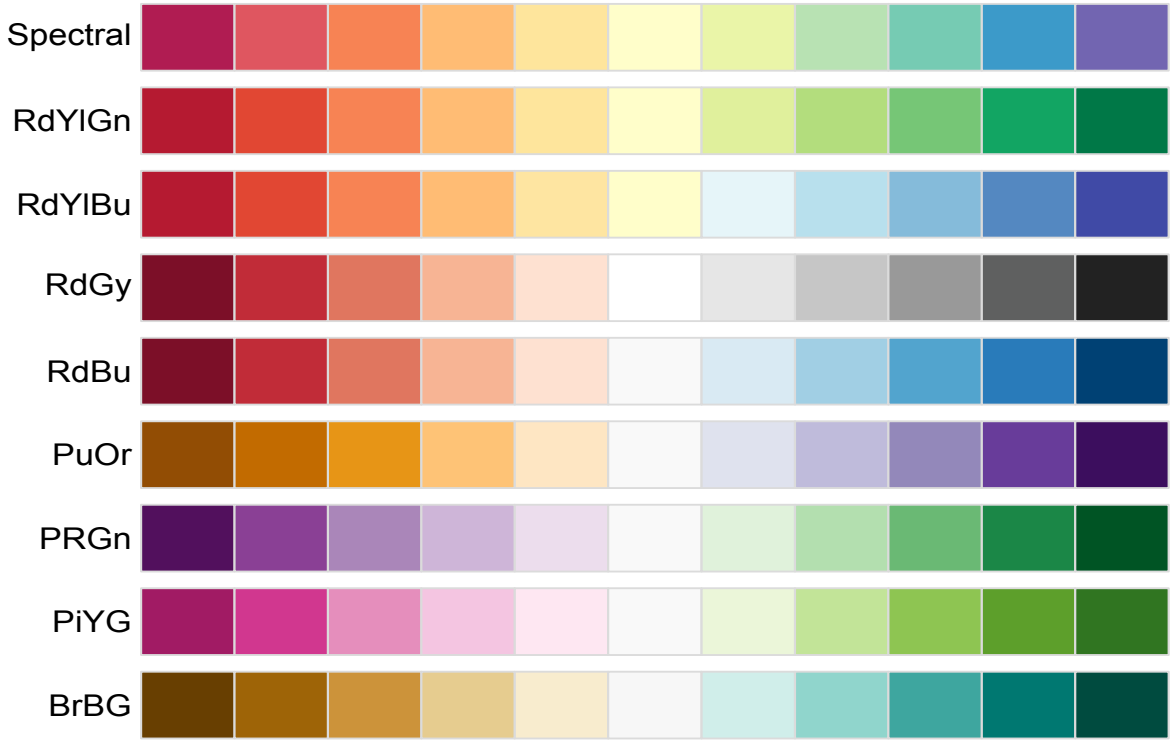
Data Type and Color

- Qualitative – Choose a **qualitative** scheme that makes it easy to distinguish between categories
- Quantitative – Choose a color scheme that implies magnitude.
  - Does the data progress from low to high? Use a **sequential** scheme where light colors are for low values
  - Do both low and high value deserve equal emphasis? Use a **diverging** scheme where light colors represent middle values

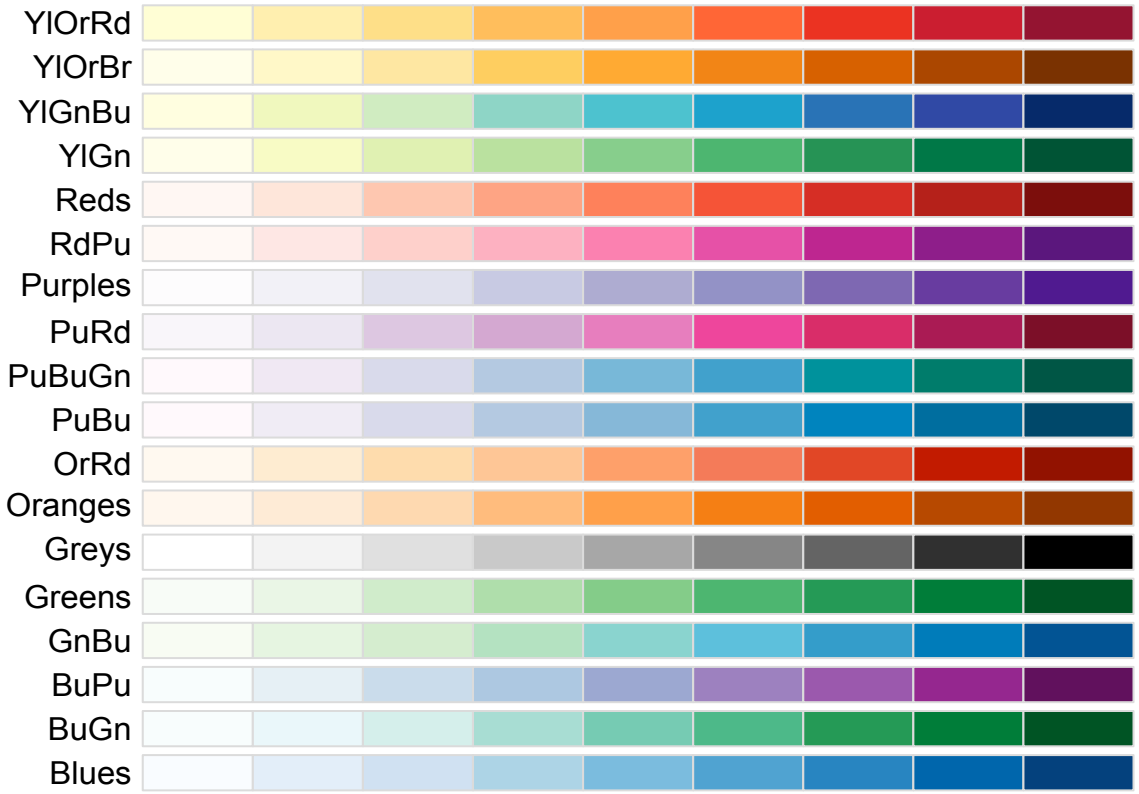
Brewer's Qualitative Palette

Cynthia Brewer, Geography PennState, tool at : http://colorbrewer2.org/

Brewer's Diverging Palette



Brewer's Sequential Palettes



Colors in R

- col="red" passed to a plotting function
- colors() : get list of 657 colors
- palette()
- rainbow(), heat.colors(), terrain.colors(), cm.colors()
- The function brewer.pal() in the package RColorBrewer provides palettes from colorbrewer2.org

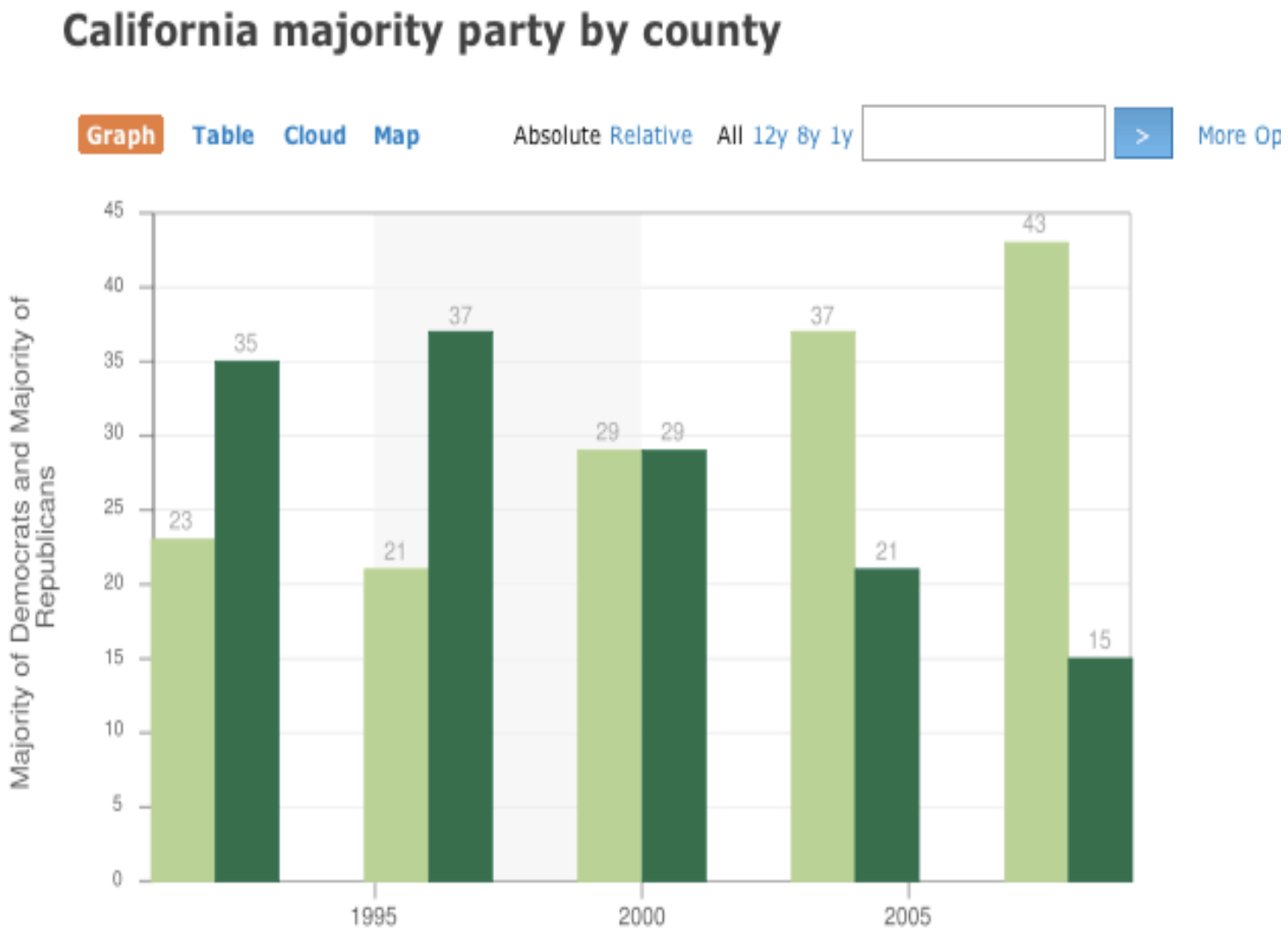
Cases

The Plotting Process

- Determine what the message is
- Help the data speak
- Plotting is an iterative process –
- An artist makes many sketches before painting the masterpiece

Case: Voter Registration Trends in California

How would you improve this plot?



Changes

- Location of tick marks under bars
- Color of bars – indicate party
- Title
- Y-axis label confusing
- X-axis label missing
- Check data for understanding of how plot is made

Data

```
Majority of Democrats, Majority of Republicans, Election Year
```

21,37,"2004"

23,35,"2008"

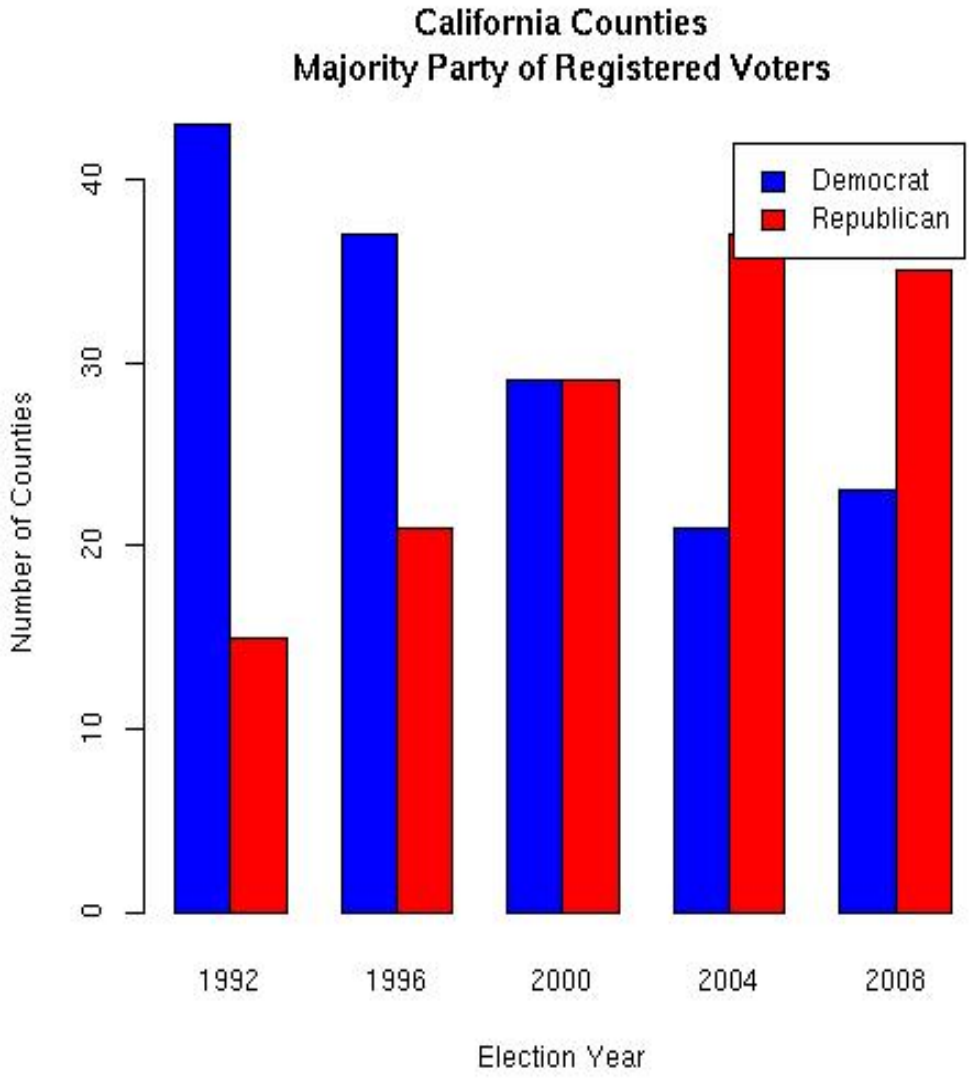
29,29,"2000"

37,21,"1996"

43,15,"1992"

Sources: California Secretary of State

http://www.sos.ca.gov/elections/ror/60day presprim/hist reg stats.pdf


California Counties
Majority Party of Registered Voters
Democrat
Republican
Number of Counties
0
10
20
30
40
1992
1996
2000
2004
2008
Election Year

What's the message?

- How party registration has changed over the past presidential elections
- More informative if we have registration figures for people not counties
- County size may be a lurking variable - small counties tend to be rural and conservative

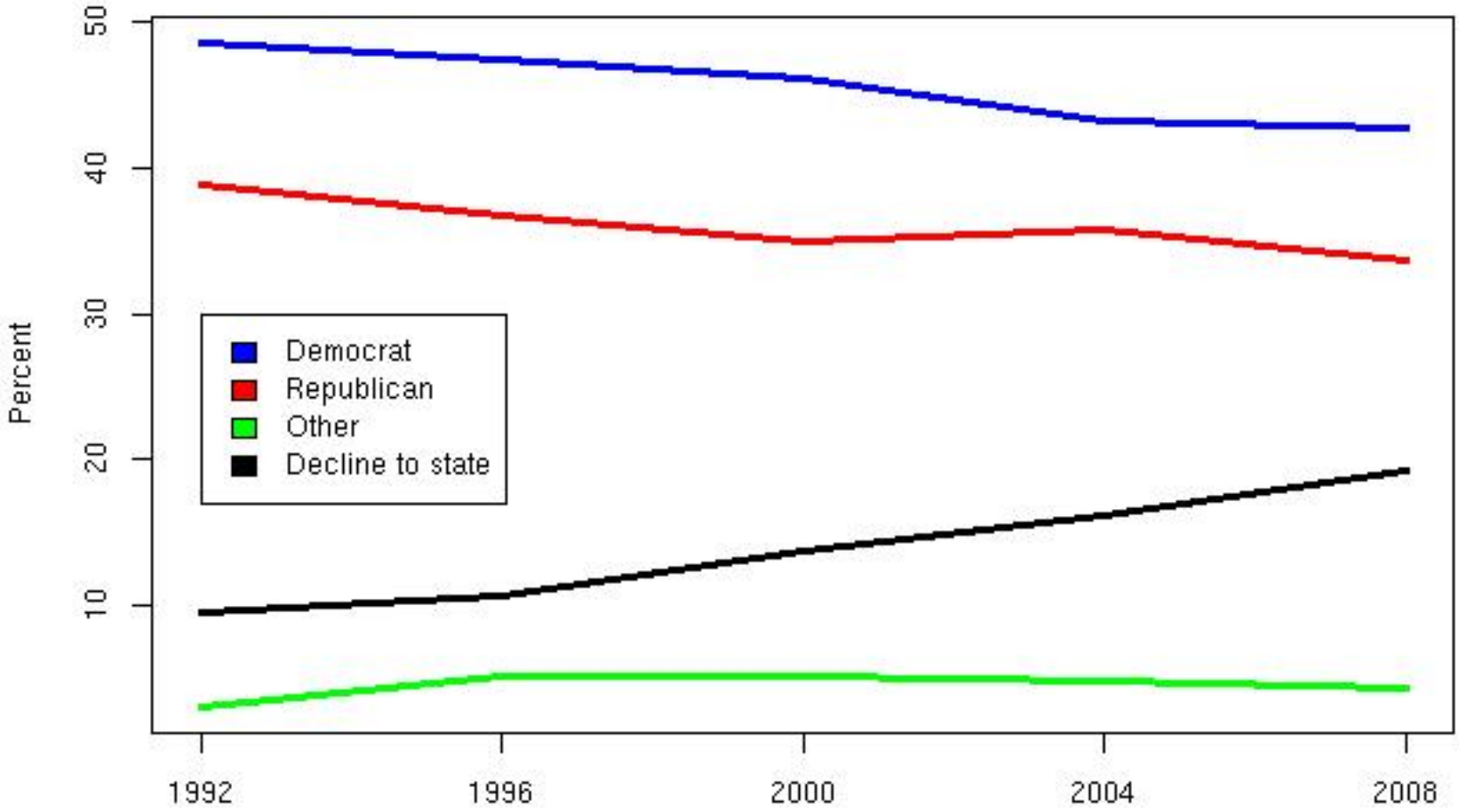
Can we make it more information rich?

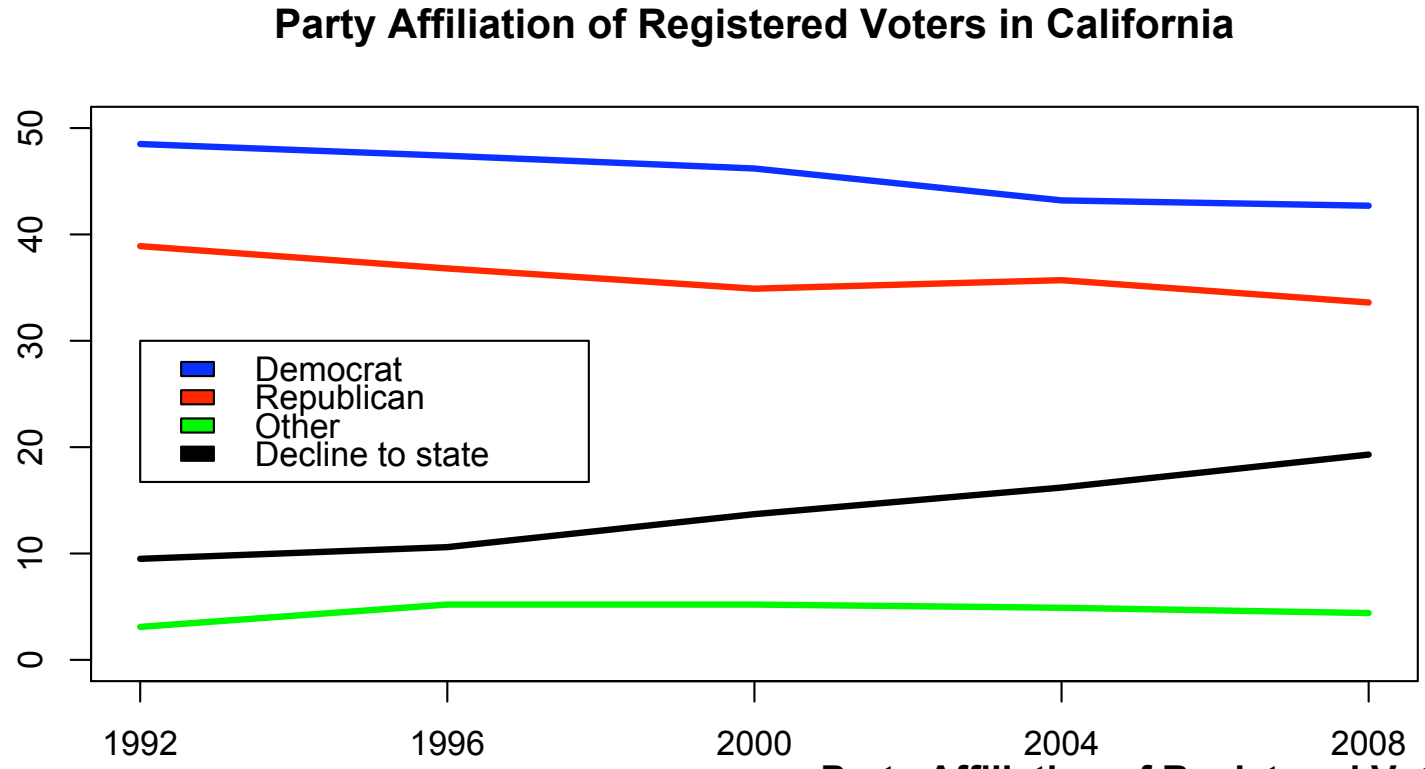
Data

year, eligible, registered, dem, rep, other, decline
1992, 20612814, 13217022,.485, .389, .031, .095
1996, 19298379, 14314658, .474, .368, .052, .106
2000, 21190865, 14676174, .462, .349, .052, .137
2004, 21843202, 14945031, .432, .357, .049, .162
2008, 22987562, 15468551, .427, .336, .044, .193

How about a line plot rather than bar chart?

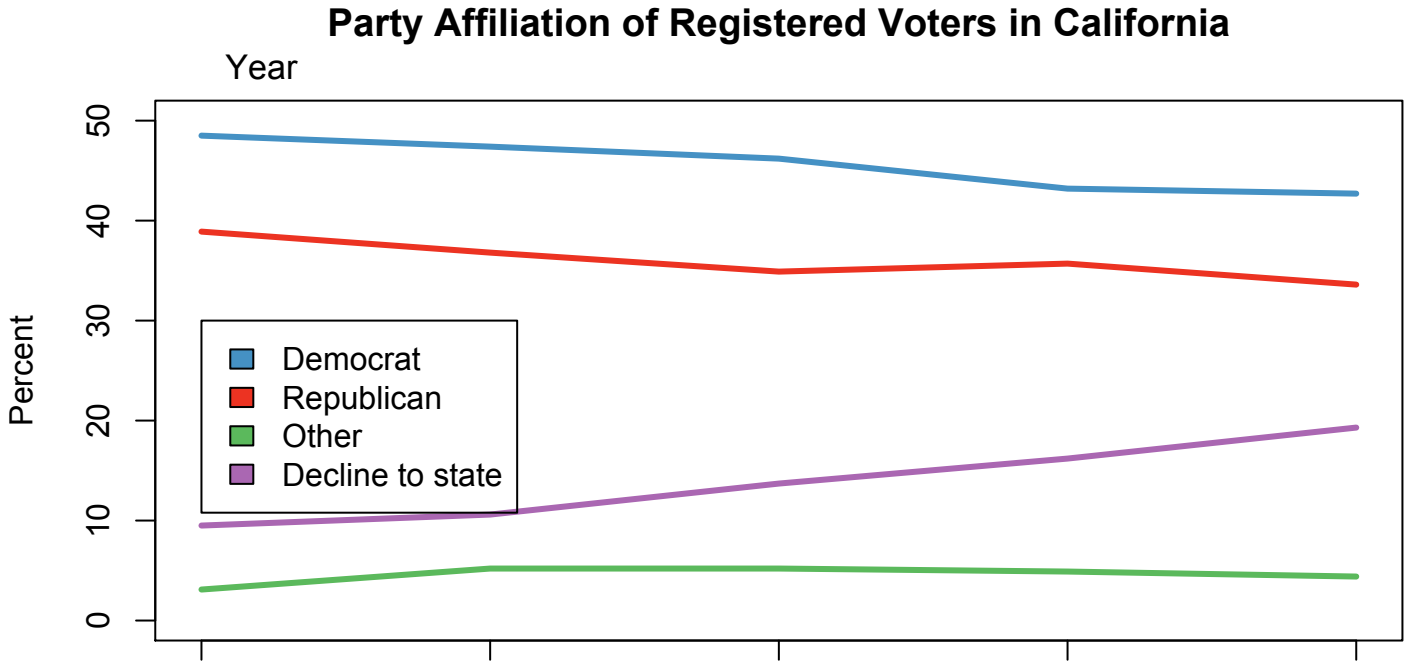
Since Other and “Decline to State” are about 25% of the 2008 registrations, leaving them out of the plot distorts the message.


Percent
10
20
30
40
50
1992
1996
2000
2004
2008
Democrat
Republican
Other
Decline to state



R default colors red, blue, green, black

Colors from Brewer's Set1 Qualitative palette



Brief look at how to use the special colors from Brewer's palettes in R

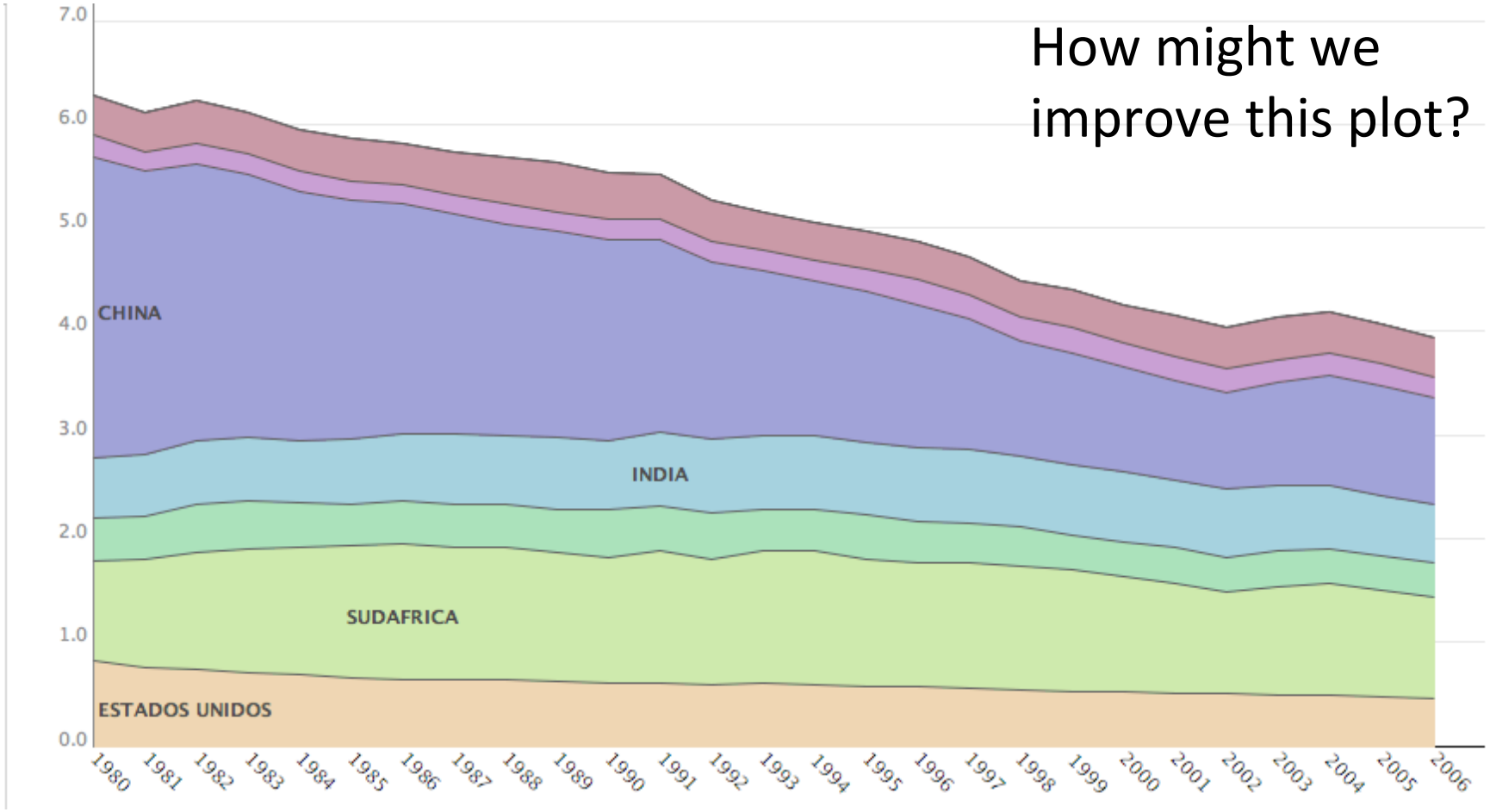
```
> library(RColorBrewer)
> colors = brewer.pal(9, "Set1")


> plot(x, y, type = "l", col = colors[1])
> colors[1]
[1] "#377EB8"
```

- R doesn't give regular English names to these colors. More later on this.

Case: CO2 emissions around the world

ManyEyes and CO_2



Changes

- Superpose rather than stack the curves so the baseline doesn’t jiggle
- Use color on the lines rather than filling polygons

Many Eyes CO_2 txt file

Uploaded by: sopecontodo
Data source: Unknown
Description:

Created at: Nov 30 2010

View as text

	Kilos de CO2 emitidos por cada dÃ³lar del PIB (PPP, 2005)	1980	1981	1982	1983	1984	1985
1	Argentina	0.38333186	0.381428156	0.406759539	0.398360537	0.394589705	0.403230413
2	Brasil	0.202990595	0.194973813	0.194266889	0.194648263	0.187318277	0.186319816
3	China	2.904045926	2.732955025	2.680960849	2.548963143	2.405590686	2.297552842
4	India	0.582845294	0.591251121	0.607423961	0.614271009	0.611892734	0.637833189
5	Mexico	0.422862904	0.425115073	0.466109144	0.452388983	0.425907729	0.394161616
6	Sudafrica	0.96297576	1.02965463	1.127518054	1.195715963	1.230043609	1.27769729
7	Estados Unidos	0.813798582	0.761893692	0.737590753	0.711116551	0.683594274	0.659206585

watch this add to topic center Visualize rate this

Read into R from the Web

```
myData  = read.table(
url("http://www.stat.berkeley.edu/users/nolan/data/
    CO2Nations.txt"),    header = TRUE, sep = "\t")
```

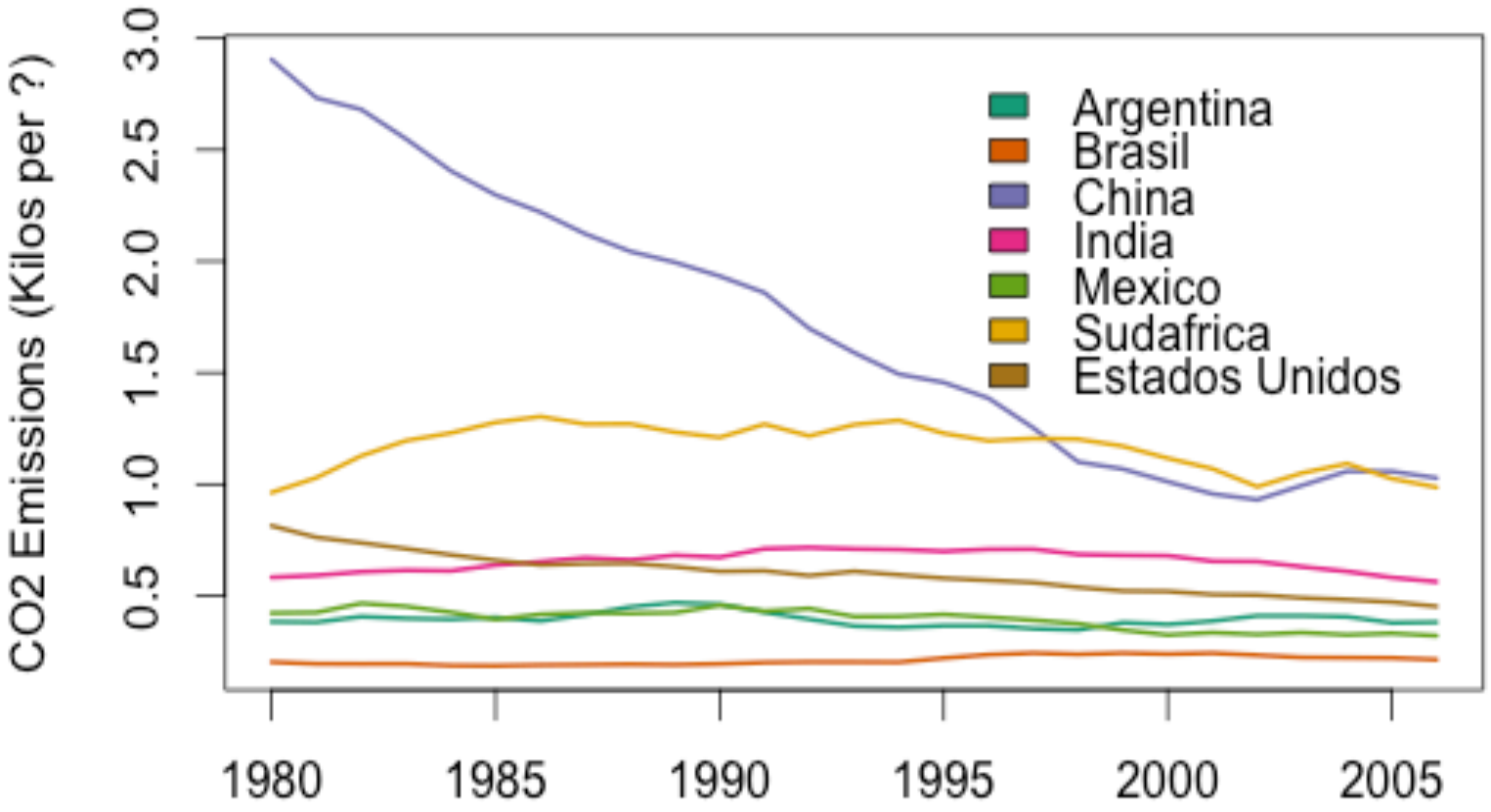
```
> head(myData)
 “Kilos….”     X1980       X1981       X1982       X1983       X1984    …
Argentina 0.3833319 0.3814282 0.4067595 0.3983605 0.3945897
Brasil       0.2029906 0.1949738 0.1942669 0.1946483 0.1873183
China      2.9040459 2.7329550 2.6809608 2.5489631 2.4055907
India       0.5828453 0.5912511 0.6074240 0.6142710 0.6118927
Mexico     0.4228629 0.4251151 0.4661091 0.4523890 0.4259077
Sudafrica 0.9629758 1.0296546 1.1275181 1.1957160 1.2300436
```

What do you notice about the data?

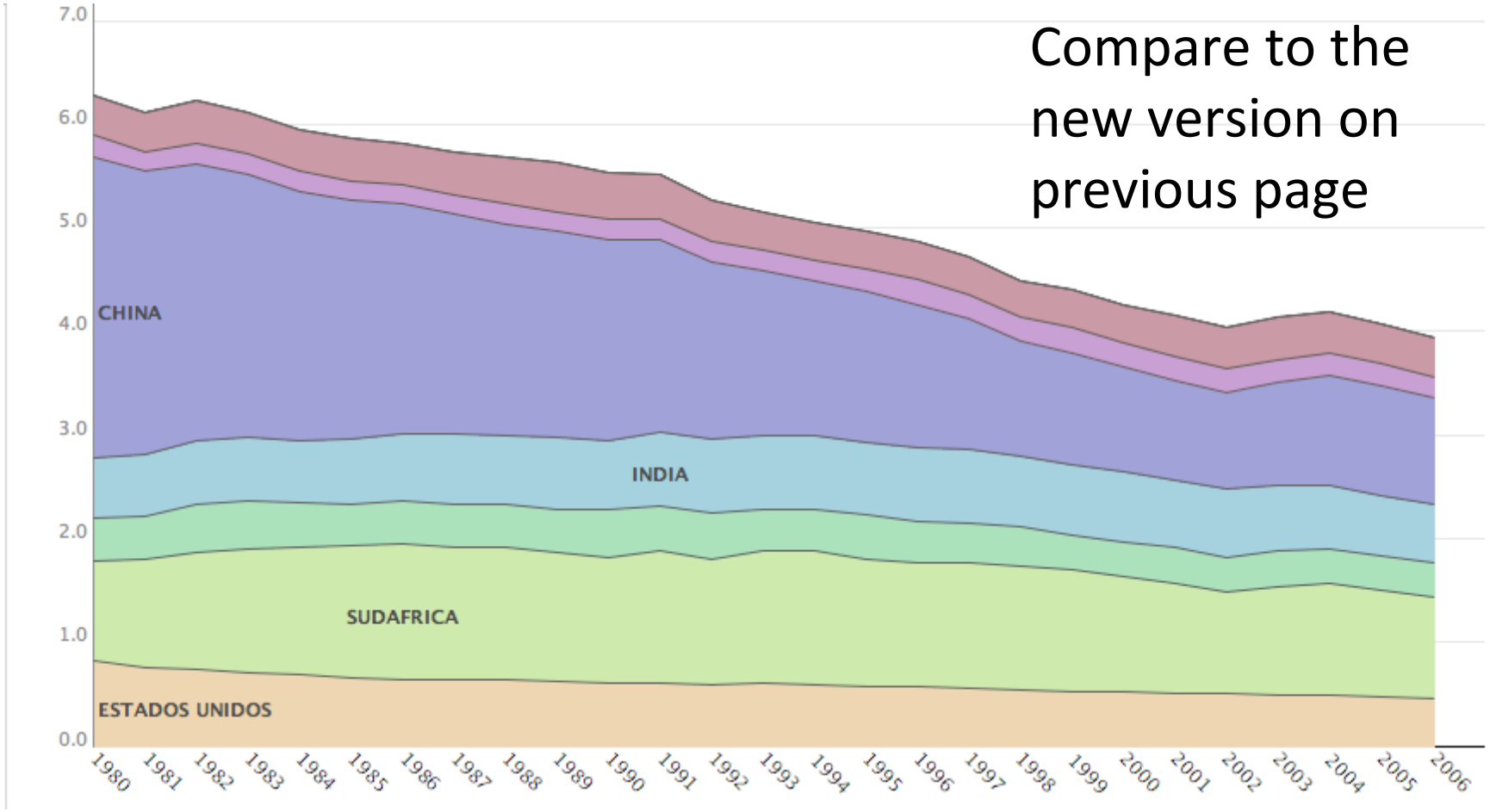
What do you notice about the data?

- The data frame has a row for every country and a column for every year
- In R, the variables are the columns of the data frame
- The variables are years, e.g., X1980. Note that R put an X in front so that the name starts with a letter.

What can you see now?



ManyEyes and CO_2

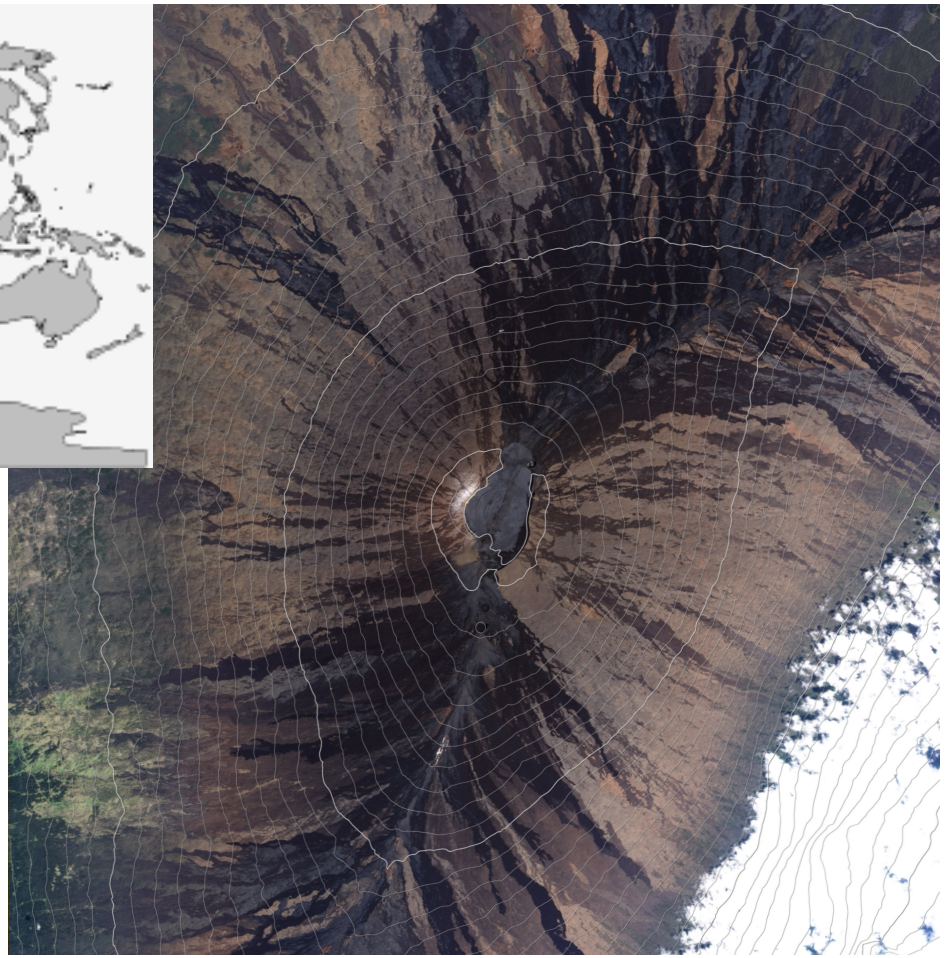


Compare to the new version on previous page

Case: CO2 levels at Mauna Loa

Time and the horizontal axis

Mauna Loa Volcano



Largest Volcano in world
4 km above sea level
Summit 17 km above base
On the Island of Hawaii

Data and photos available from Scripps Institute and NOAA

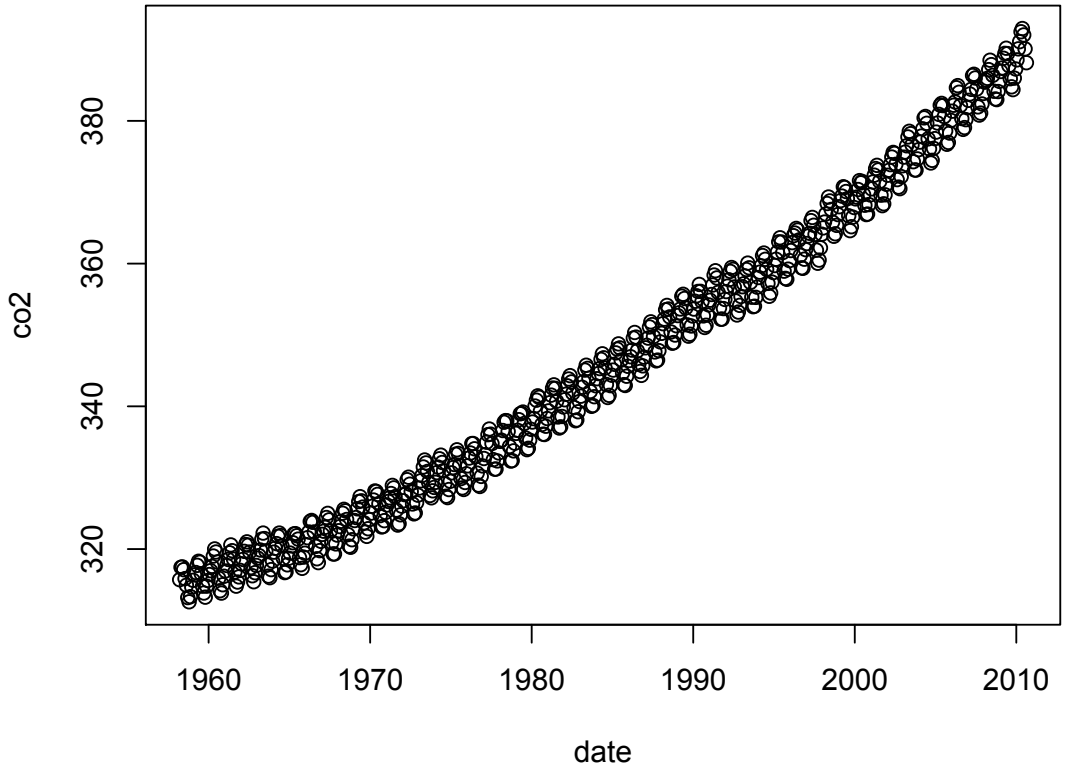
Mauna Loa Observatory

- Far from any continent, the air sampled is a good average for the central pacific.
- Being high, it is above the inversion layer where local effects are present.
- Measurements of atmospheric CO_2 since 1958 – longest continuous record

Atmospheric Carbon Dioxide

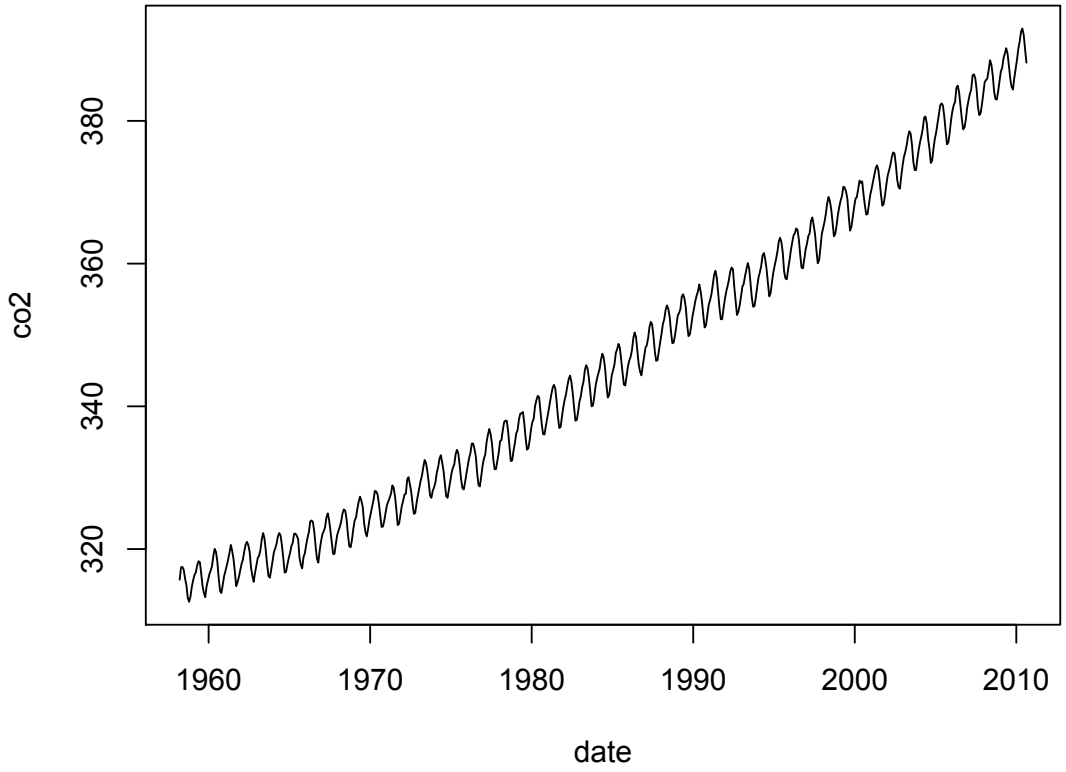
- The increasing amount of CO_2 in the atmosphere from the burning of fossil fuels has become a serious environmental concern.
- Upper safety limit for atmospheric CO_2 is 350 parts per million
- Does a rise in CO_2 lead to a rise in world temperatures?

Time Series – Pairs: (time, CO_2)

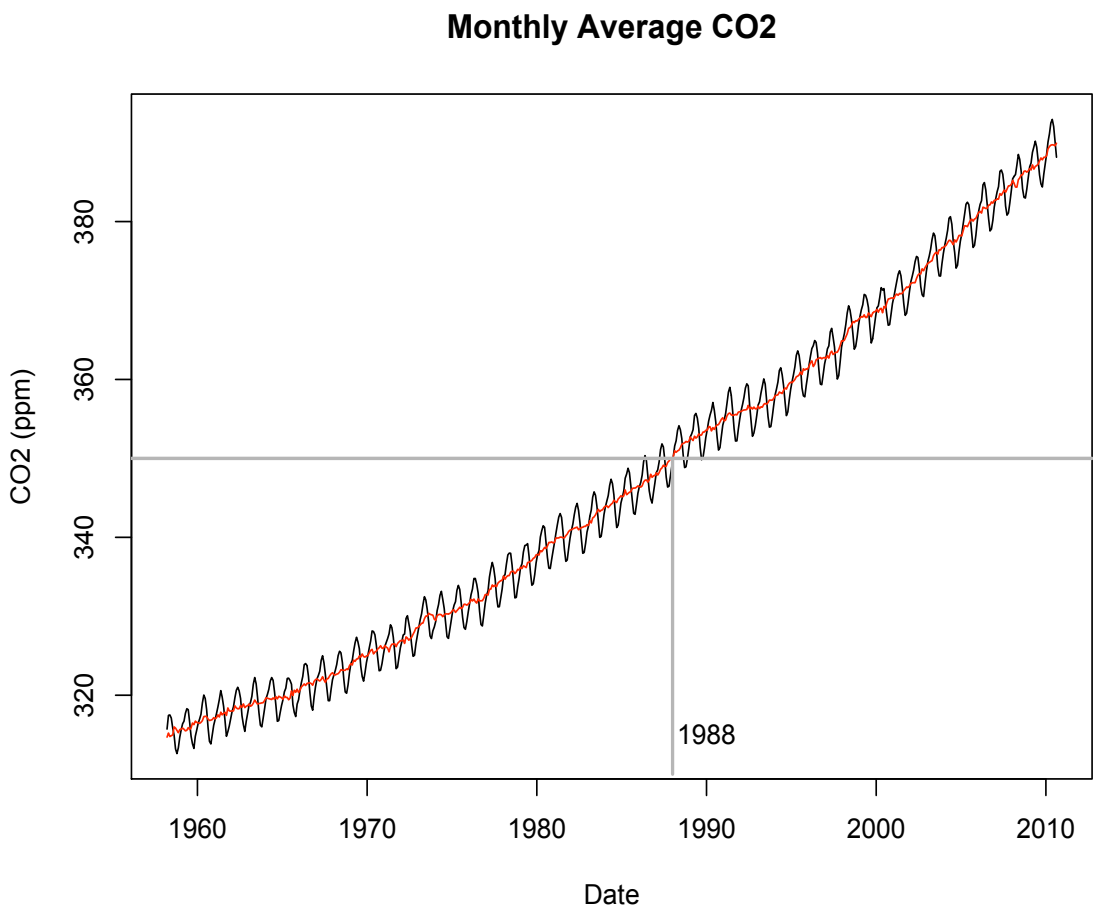


Points are typically not the best way to plot time series

Connect the measurements with line segments



Seasonality vs the long-term Trend



Aspect Ratio

- The height/width of the data region was selected to be about 1 so that the trend line is at about 45 degrees.
- The banking to 45 degrees let's us see that the curve is convex
- This means that the rate of increase of CO_2 is increasing through time

Global Warming

- 1981 US Senate convened scientists for testimony on global warning
- Senator Al Gore said that the Mauna Loa data clearly demonstrated increases in CO_2
- Pewitt (witness for the DOE) said that the graph was misleading because it doesn't include 0

Chartology

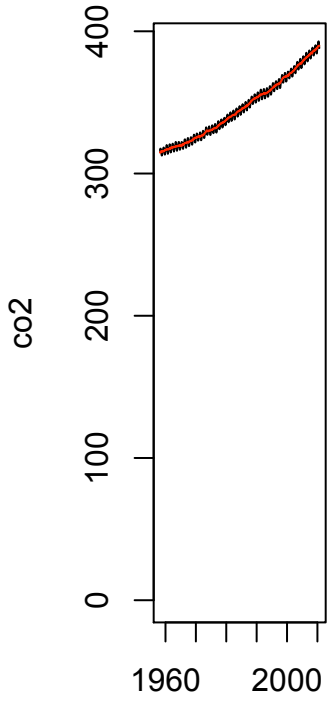
Pewitt took issue with the graph, saying

"It is a clever piece of chartology" because it can be read the wrong way.

He continued, "It is intellectually just exactly correct. It displays 315 going to 336, but it appears to be going from 0 to very large amounts."

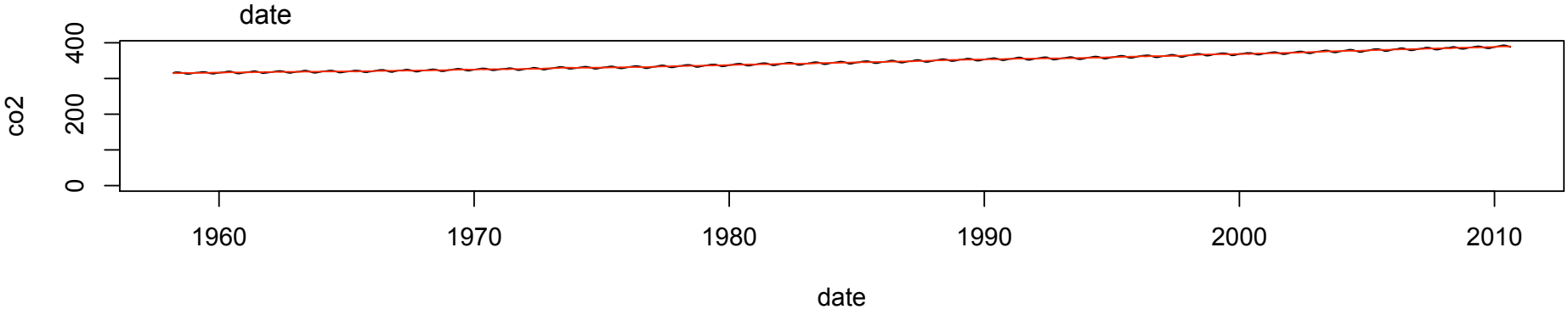
Steven Schneider (*Global Warming*) called Pewitt's objection "double talk"

Including 0 & The Aspect Ratio



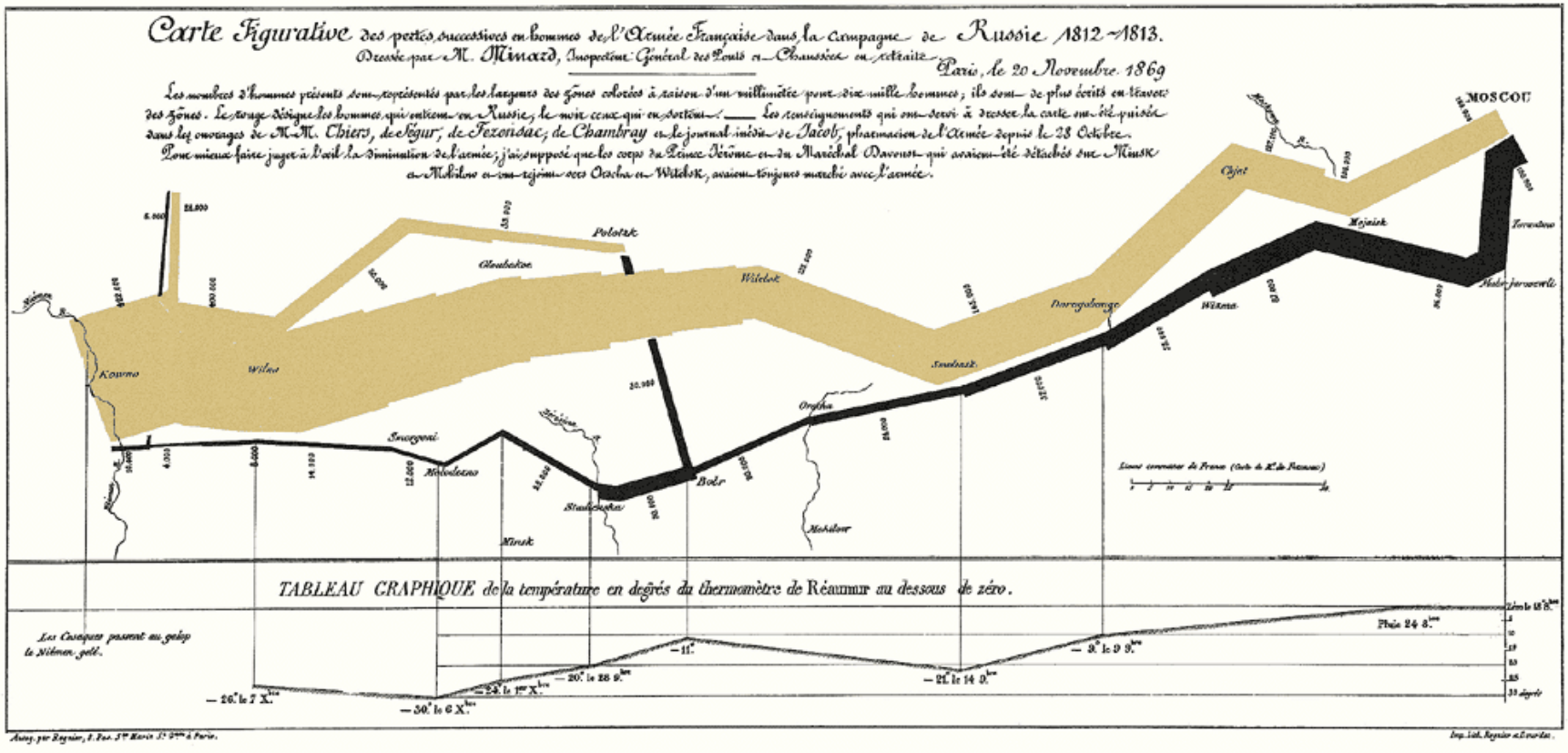
When we include 0, if we also bank at 45 degrees, the plot must be tall and narrow.
With this plot it's hard to see any other features.
There is also a lot of empty space.

To fill the space with data, we need to stretch the data region to be wide and short.
Now, it's hard to see the most important feature because the banking is nearly 0.



Case: Napoleon’s March

Minard's Napoleon's March



Minard Map

- Size of Army – thickness of the band
- See the effect of individual battles, e.g. the crossing of Berezina
- Clear, effective summary
- “seeming to defy the pen of the historian by its brutal eloquence,” E.J. Marey

Minard's Data

- Size of Army
- Date
- Location
  - Latitude
  - Longitude
- Temperature
- Direction (advance/retreat)

Minard's Napoleon's March