Graphics I

Outline

- Review plot types
  - Univariate
  - Bivariate
  - Multivariate
- Design
  - Best practice
  - Pitfalls
  - Critique
  - Perception & color
- Graphics Models in R
  - Base
  - Grammar of Graphics
- Information visualization vs Statistical graphics

Types of Plots

Know your data types

The appropriate graphical techniques depend on the kind of data that you are working with

- Quantitative
  - continuous – e.g., height, weight
  - discrete – numeric data with few values, e.g., number of children in family
- Qualitative
  - ordered – categories with an order but no meaningful distance between, e.g., number of stars for a movie rating
  - nominal - categories have no meaningful order, e.g., gender

One Variable

Case: Infant Health

```
load(url("http://www.stat.berkeley.edu/
users/nolan/data/KaiserBabies.rda"))
```

Kaiser Study

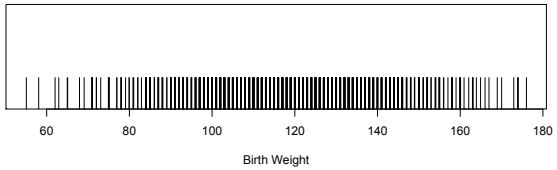
- Oakland Kaiser mothers
- 1960s
- Measure the babies weight (in ounces) at birth
- All babies:
  - Male
  - Single births (no twins, etc.)
  - Survived 28 days

Information collected on mother's and their babies

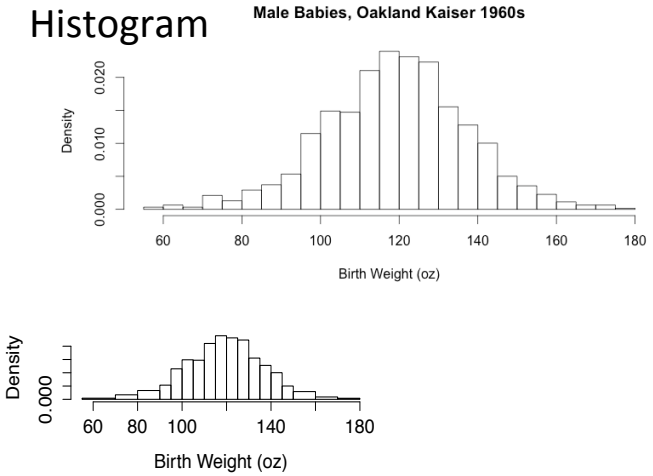
- Birth weight (ounces)
- Gestation (weeks)
- Parity - total number of previous pregnancies
- Mother's height and weight
- Mother's smoking status
- Mother's age, race, education level, income
- And more…

Rug plot & Strip chart

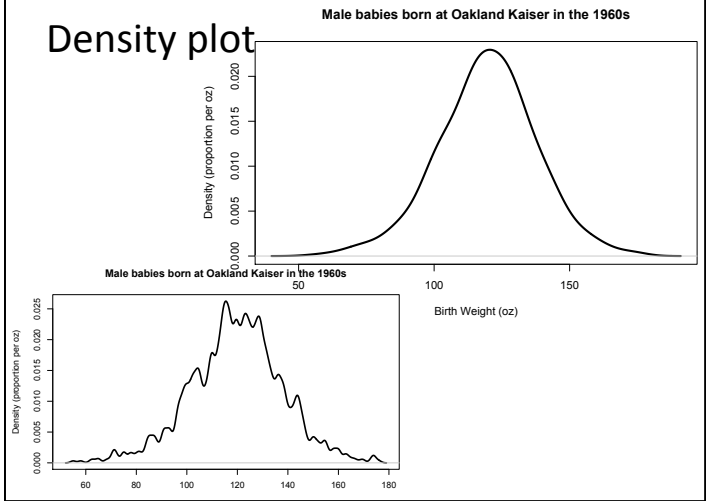
Each baby's weight is represented as a tickmark. The thicker lines are from multiple babies with similar weights. I added a little random noise to the weights to keep them from falling on top of each other.



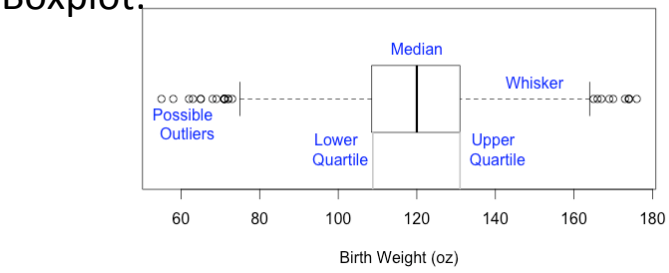
Histogram

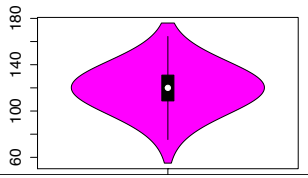


Density plot



Boxplot:





Violin Plot

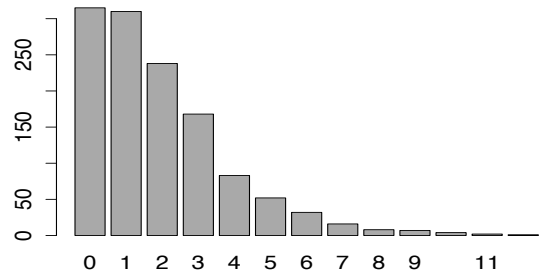
Parity: Number of siblings

- This quantitative variable is different from birth weight – there are only a few possible values, i.e., it's not possible to have 2.3 siblings, and it's highly unlikely to have 17

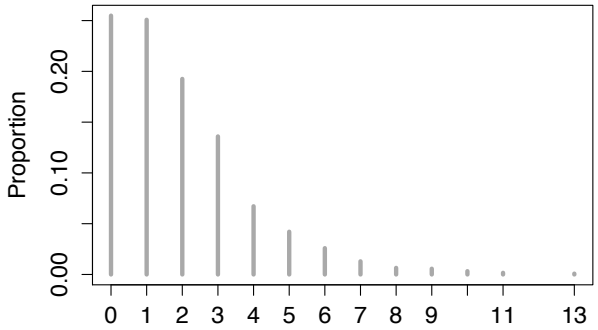
```
> table(infants$parity)
```

0	1	2	3	4	5	6	7	8	9	10	11	13
315	310	238	168	83	52	32	16	8	7	4	2	1

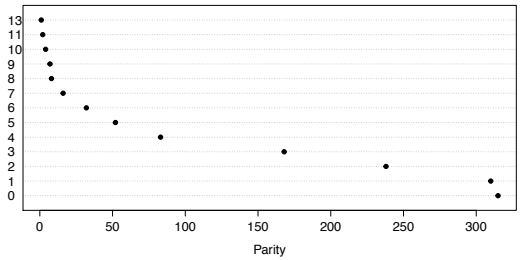
Number of Siblings



Alternative – bar width has no meaning



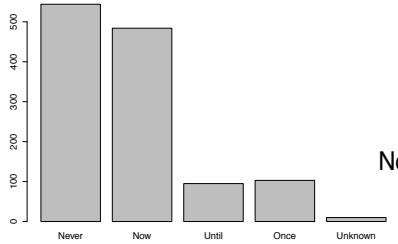
Dot chart



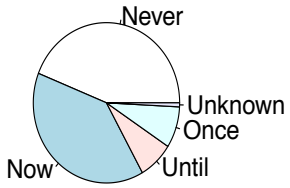
Qualitative Variables

Smoking Status - Categorical

Bar chart



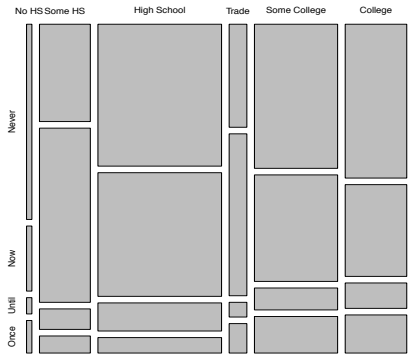
Pie chart



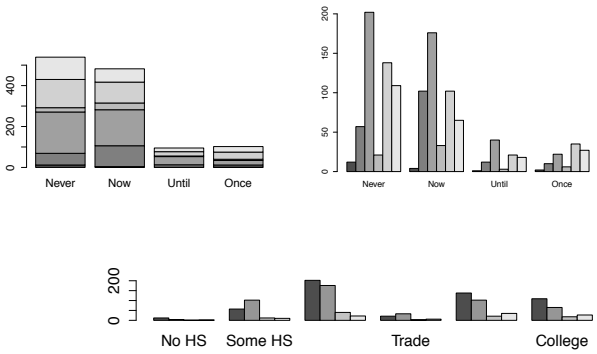
ANGLES can be hard to compare

Two Variables

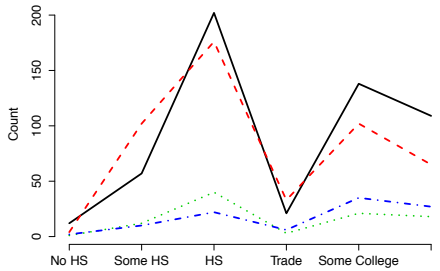
Two Qualitative – Mosaic plot



Barplots: side-by-side & stacked



Line plots



San Francisco Chronicle listings

load(url("http://www.stat.berkeley.edu/users/nolan/data/Projects/SFHousing.rda"))

Data

- Record: house sold in a particular time period
- Over 200,000 houses
- Subset to a dozen cities in the East Bay – about 25,000 houses

Variables:

- City
- County
- Price
- # bedrooms
- Lot square footage
- and 10 more

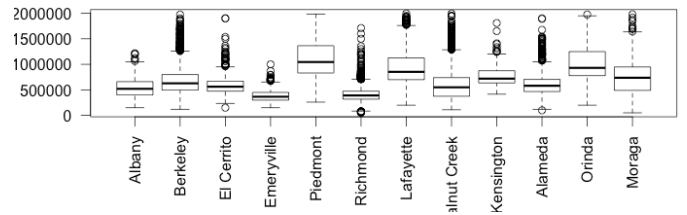
Relationship between city and sale price

Data types:

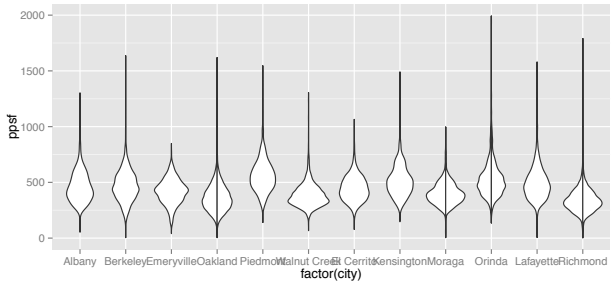
City - factor

Sale price - numeric

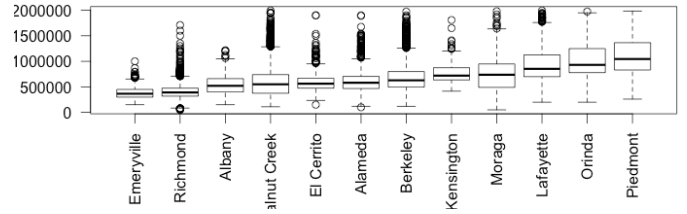
Side-by-Side Boxplots



Side-by-side Violin plots



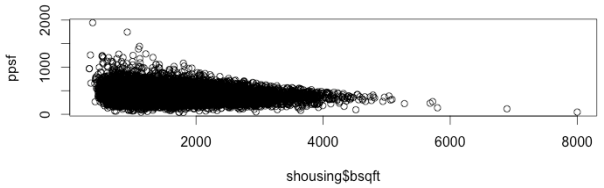
Cities ordered by median price

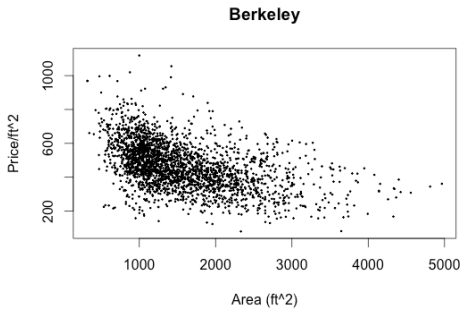


Relationship between price per square foot and total square foot

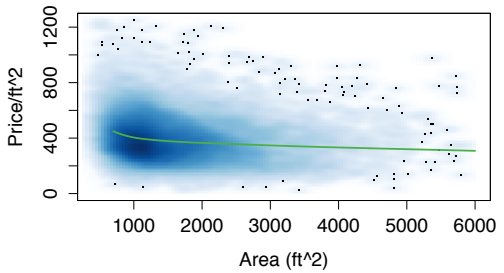
Both are quantitative

WHAT's Wrong with this plot?





Smooth scatter plot

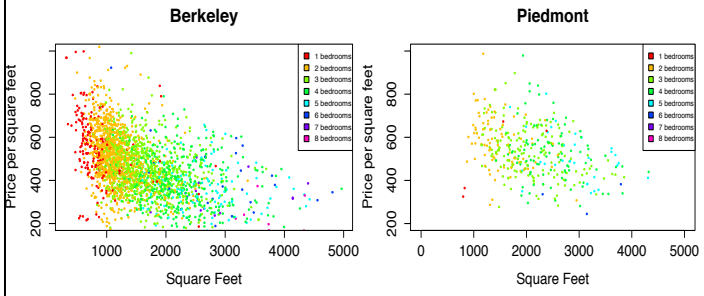


Summary of graph relationships between two variables

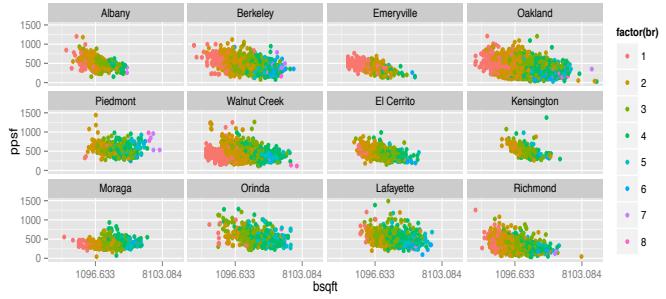
- Two Qualitative variables
  - mosaicplot, side-by-side barplots, line plots
- One Quantitative and one Qualitative
  - Boxplots, dotcharts, multiple density plots, violin plots
- Two Quantitative variables
  - Scatter plot, line plot

Relationships between more than 2 variables

- Qualitative information can be conveyed in plots through color, plotting symbol, juxtaposed panels
- The following plot uses information from 4 variables: city, number of bedrooms, lot size (sq ft), and price per square ft



Bsqft, PPSF, Bedrooms, City



Scatterplots of all pairs of variables

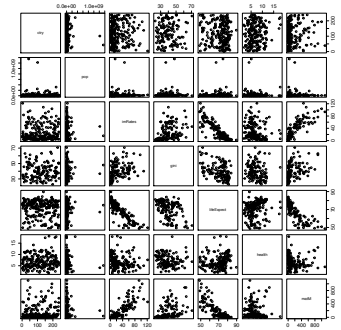
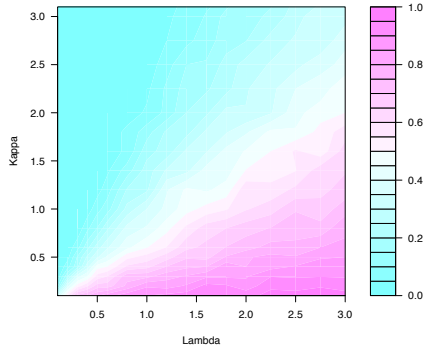
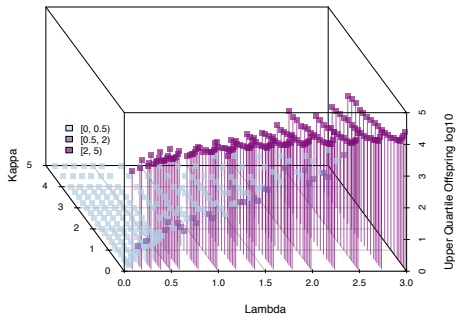


Image Map with 3-numeric variables



3D Scatterplot

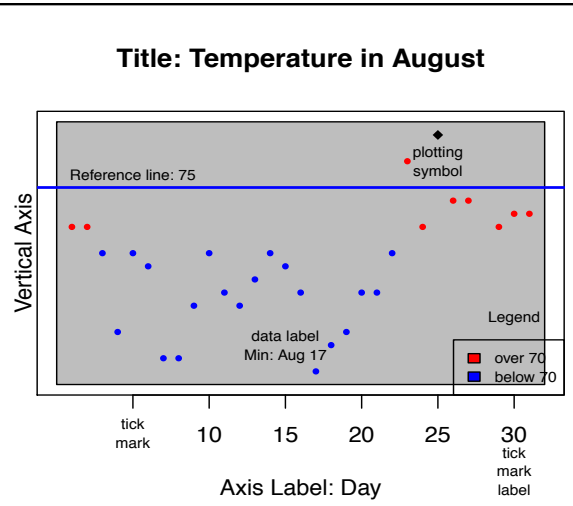


Best Practices

Outline

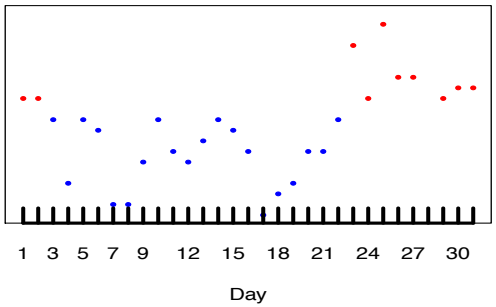
- Vocabulary
- 3 Properties of good graph construction
  - Data stand out
  - Facilitate comparison
  - Information rich
- Perception
- Case studies

Vocabulary

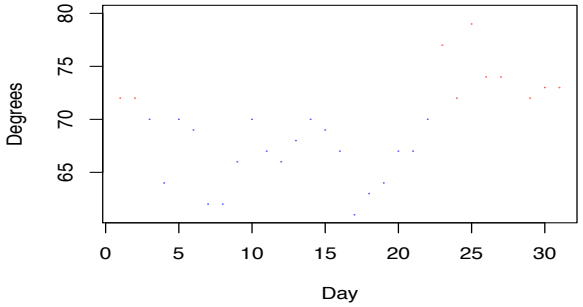


1. Make the Data Stand Out

Avoid having other graph elements interfere with data

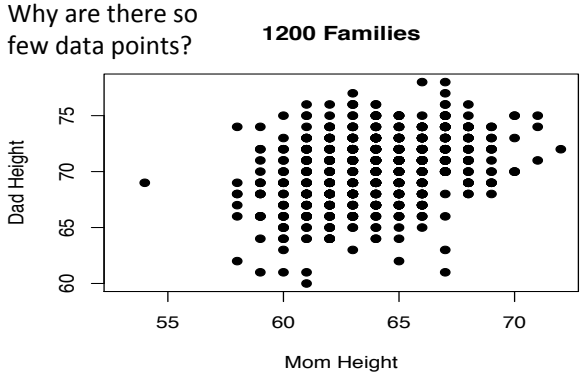


Use visually prominent symbols



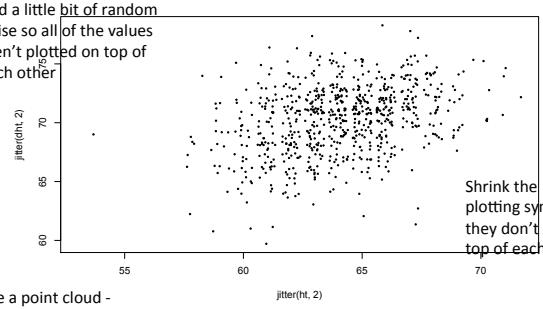
Avoid over-plotting

Why are there so few data points?



One way to avoid over plotting: Jitter the values

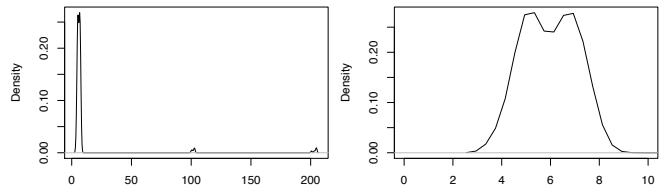
Add a little bit of random noise so all of the values aren't plotted on top of each other



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

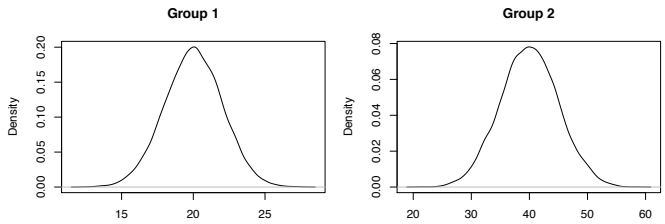
- Include all or nearly all of the data
- Fill data region
- Origin need not be on the scale
- Choose a scale that improves resolution (to be continued)

Eliminate superfluous material

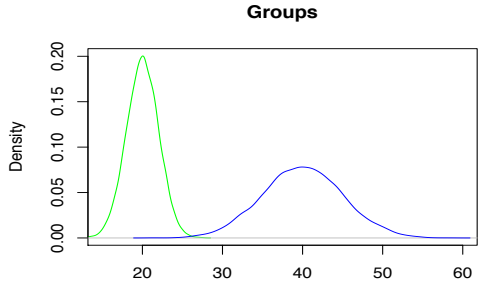
- Chart junk – stuff that adds no meaning, e.g. butterflies on top of barplots, background images
- Extra tick marks and grid lines
- Unnecessary text and arrows
- Decimal places beyond the measurement error or the level of difference

2. Facilitate Comparisons

Put Juxtaposed plots on same scale



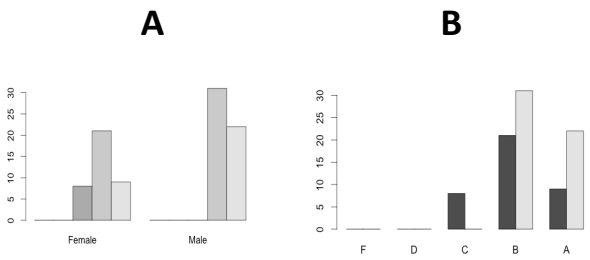
Make it easy to distinguish elements of *superposed* plots (e.g. color)



Choosing the Scale

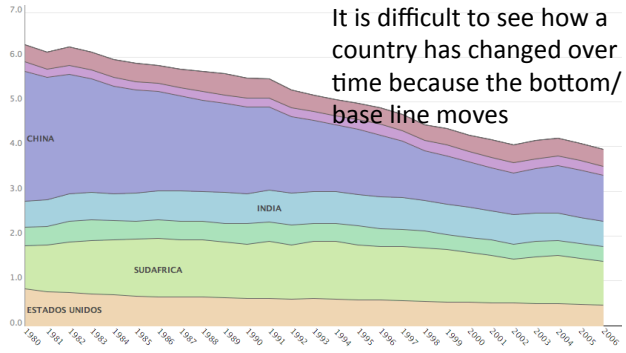
- Keep scales on x and y axes the same for both plots to facilitate the comparison
- Zoom in to focus on the region that contains the bulk of the data
- These two principles may go counter to one another
- Keep the scale the same throughout the plot (i.e., don't change it mid-axis)

Emphasizes the important difference



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

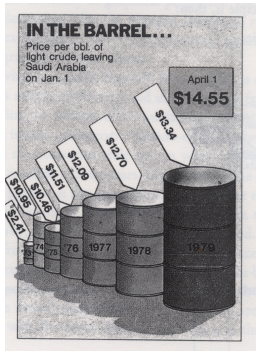
Avoid Jiggling the baseline



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



Perception

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

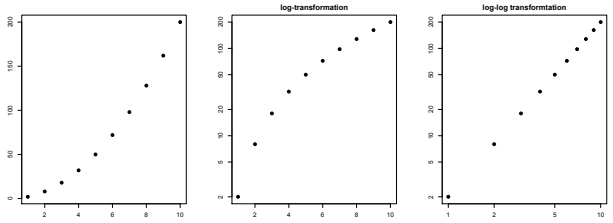
Banking: Aspect Ratio

- 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

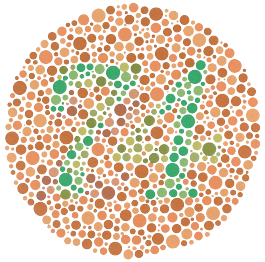
Bar plot vs Pie chart

- 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.

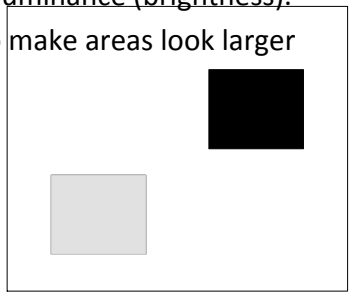


Colorfulness

- 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



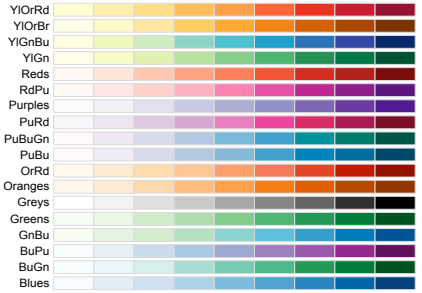
Data Type and Color

- 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 Palettes

Brewer's Diverging Palettes

Brewer's Sequential Palettes



3. Add Information

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
- 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

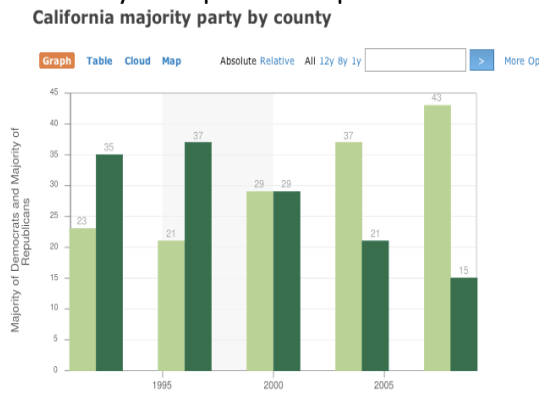
Cases

The Plotting Process

- Determine what's the message
- 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

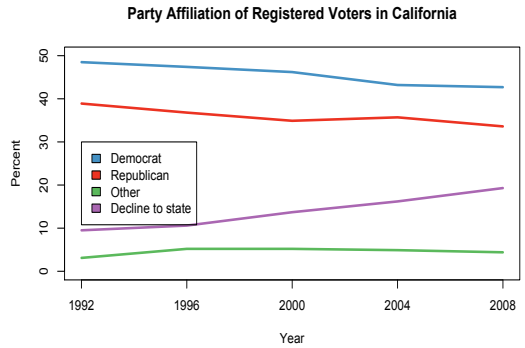
- Location of tick marks under bars
- Color of bars – indicate party
- Title
- Y-axis label confusing
- X-axis label missing
- Check data for understanding of how plot is made

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?

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

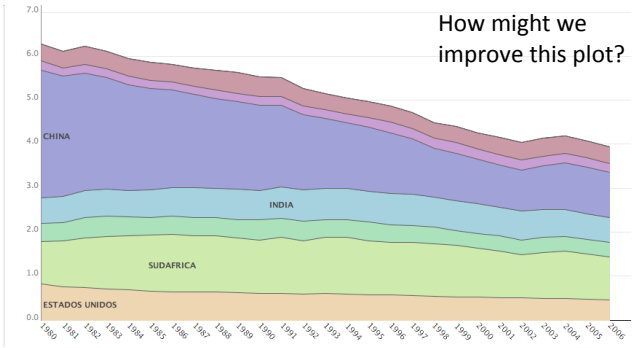


Colors from Brewer's Set1 Qualitative palette

Case: CO2 emissions around the world

ManyEyes and CO_2

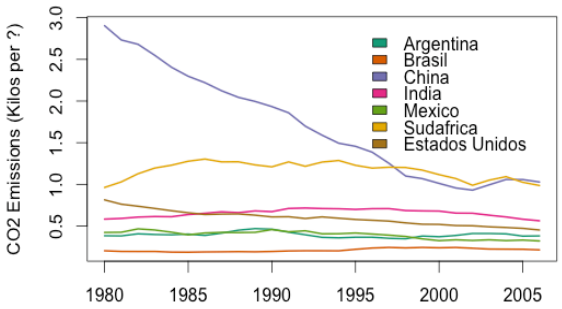
How might we improve this plot?



Changes

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

What can you see now?

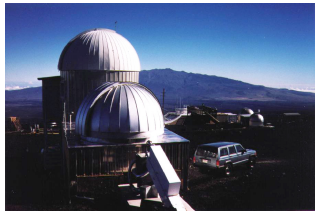


Case: CO2 levels at Mauna Loa

Time and the horizontal axis

Mauna Loa Observatory

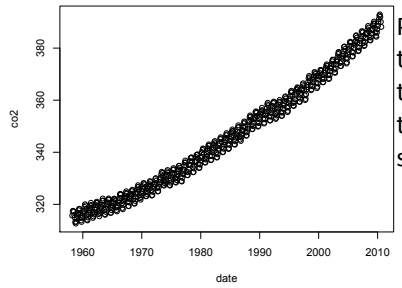
- Far from any continent, the air sampled is a good average for the central pacific.
- Being high, it is above the inversion layer where local effects are present.
- Measurements of atmospheric CO_2 since 1958 – longest continuous record



Atmospheric Carbon Dioxide

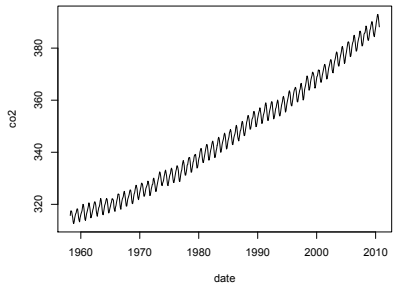
- The increasing amount of CO_2 in the atmosphere from the burning of fossil fuels has become a serious environmental concern.
- Upper safety limit for atmospheric CO_2 is 350 parts per million
- Does a rise in CO_2 lead to a rise in world temperatures?

Time Series – Pairs: (time, CO_2)

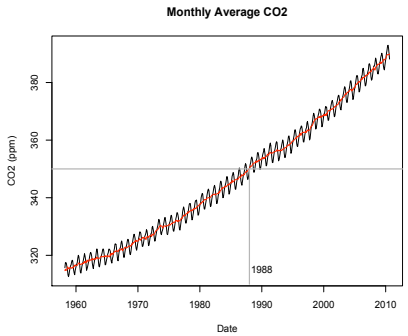


Points are typically not the best way to plot time series

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 scientist 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

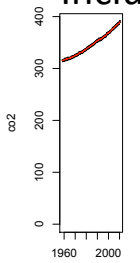
Pewitt took issue with the graph, saying

"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."

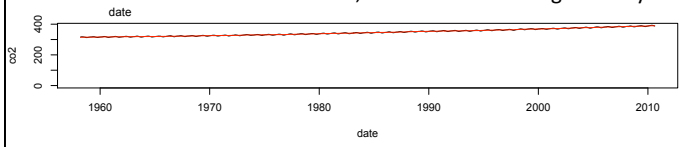
Steven Schneider (*Global Warming*) called Pewitt's objection "double talk"

Including 0 & The Aspect Ratio

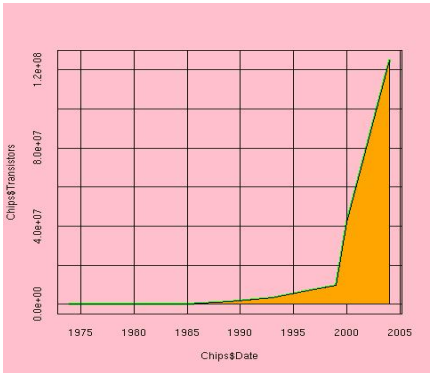


When we include 0 and bank at 45 degrees, then the plot must be tall and narrow.
With this plot it's hard to see any other features.
There is also a lot of empty space.

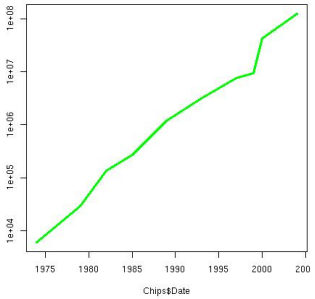
To fill the space with data, we need to stretch the data region to be wide and short.
Now, it's hard to see the most important feature, the curvature, because the banking is nearly 0.



What do you think of this plot?



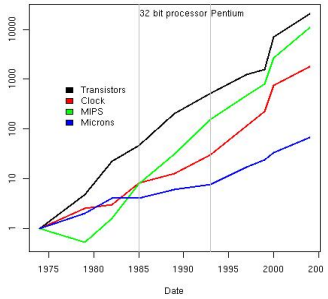
FIND 5 things that you would change



Eliminate:

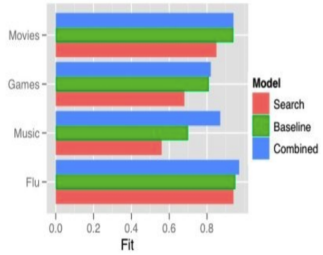
- background color,
- Grid lines,
- Polygon filling

Take logs

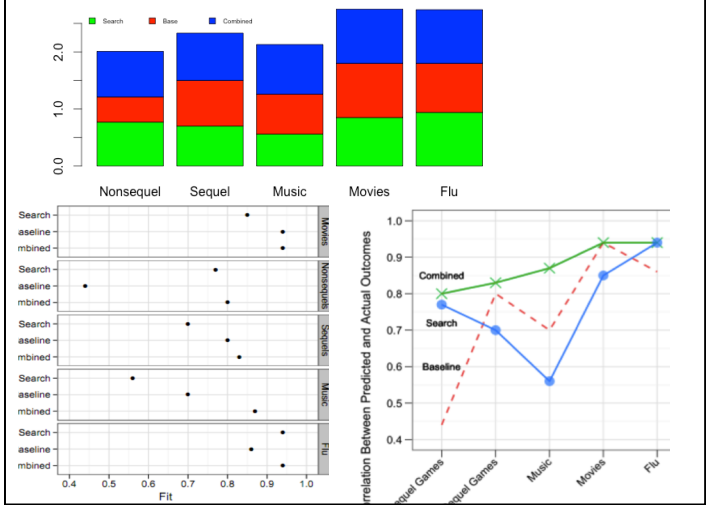


- Examine relative growth so we can add more variables
- Add legend for different variables
- Add reference lines for important dates

What alternative plot would better enable comparisons?



A. Stacked Barchart
B. Dot Chart
C. Pie Chart
D. Line plot
E. Mosaic plot



Graphics Best Practices - Recap

Data Stand Out

- Avoid having other graph elements interfere with data
- Use visually prominent symbols
- Avoid over-plotting
- Choose appropriate axis scales
- Eliminate superfluous material

Facilitate Comparisons

- Juxtapose or Superpose plots (use same scale)
- Don't change scale mid-axis
- Use only one scale on one axis
- Use color
- Emphasize the important difference
- Avoid jiggling the baseline
- Care with the visual metaphor (linear, area, volume

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

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

Graphics Models in R

R's graphics model

- There are two models in R –
- The painter's model, which is the base graphics
- The object-based model, which follows Wilkinson's Grammar of Graphics

R's base graphics model

- High-level function plot()
  - Starts a new plot on a new page
  - creates a complete plot
- Low-level functions add more output to the current plot, if necessary
  - lines and line segments
  - points, text
  - legend

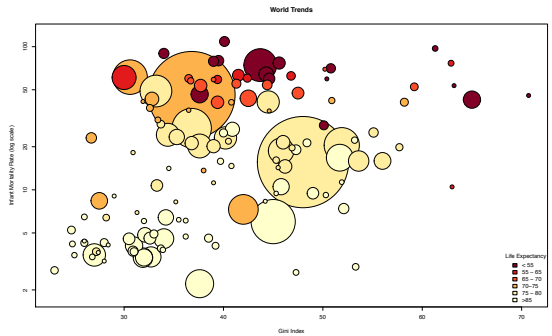
Review Plotting Functions

- `hist()`
- `boxplot()`
- `dotchart()`
- `plot()`
- `barchart()`
- `mosaicplot()`
- `abline()`
- `curve()`
- `points()`
- `lines()`
- `polygon()`
- `text()` add text

Review Plot Arguments

`?plot.default`

- `type = "l"` "p" for points, "l" for lines, "n" for nothing
- `ylim = c(0, 1)` the range for the scale of the axis
- `xlab = "x axis label"`
- `main = "plot title"`
- `col =` vector of colors
- `log = "y"` use log scale on y axis, can be "x" or "xy"
- `lwd = 2` thickness of line
- `pch = 19` plotting character – check other numbers
- `cex = 0.5` character magnification
- `lty = 2` type of line – check other numbers
- `las = 1` 0,1,2, or 3 style of tick mark labels




```
plot(x = fbDF$gini, y = fbDF$im,
     type = "n", log = "y",
     xlab = "Gini Index",
     ylab = "Infant Mortality Rate (log scale)",
     main = "World Trends" )

symbols(x = fbDF$gini, y = fbDF$imRates,
        bg = myColors[fbDF$cutLE], add = TRUE,
       circles = pmax(sqrt(fbDF$pop)/8000, 0.2),
        inches = FALSE)

legend("bottomright", title = "Life Expectancy",
       legend = lifeLabels),
       bty = "n", fill = myColors)
```

ggplot2 model

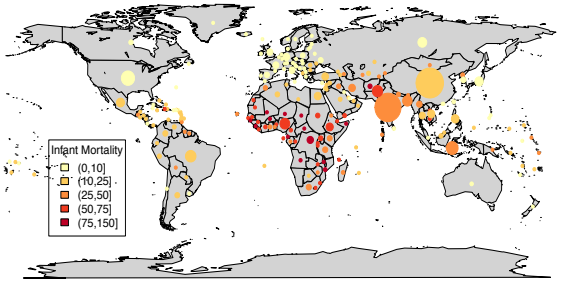
- Functions produce plot components (aka layers)
- Combine components in different ways with the + operator to produce a variety of plots

ggplot2 model - steps

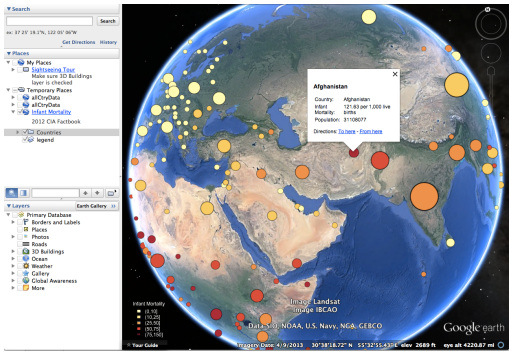
- Define the data you want to plot and create an empty plot object with `ggplot()`
- Specify what graphcis shapes, aka, geoms, to view the data, e.g., plotting symbols, lines. Add these to the plot with `geom_point` or `geom_line`
- Specify features/aestheitcs used to represent the data, e.g., *x, y* locations, color, symbol size

```
p = ggplot(fbDF)

p +  geom_point(aes(x = gini, y = imRates,
                color = cutLE, size = popS)) +

  scale_y_continuous(trans = "log",
   name = "Infant Mortality Rate (log scale)") +

  scale_x_continuous(name = "Gini Index") +

  scale_color_manual(values = myColors,
                     name = "Life Expectancy") +

  scale_size(name = "Population") +
  ggtitle("World Trends")
```

With geographic information we can make maps



We can use apps like Google Earth to display our data



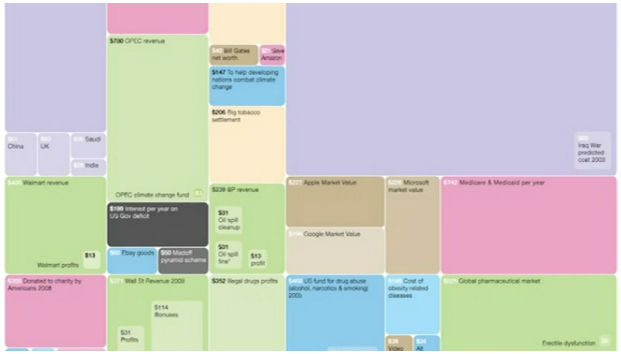
Information Visualization vs Statistical Graphics

Info-Viz vs Statistical Graphics

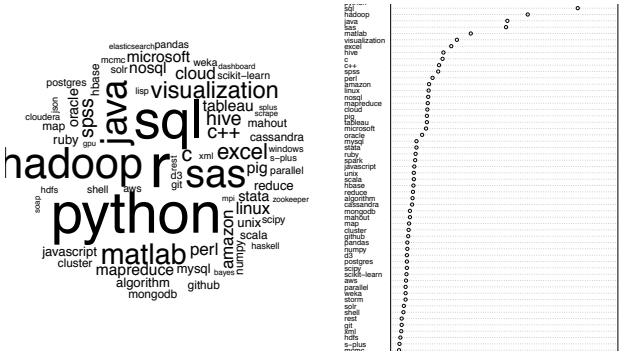
- Information visualizations tend to be grabby and visually striking
  - Dramatize the problem with a unique and visually-appealing image that draws casual viewer in deeper
- Statistical graphics try to reveal patterns and discrepancies and make comparisons
  - Careful to make judicious use of all aspects, often for viewers already interested in the problem

From Gelman and Unwin's article (2013)

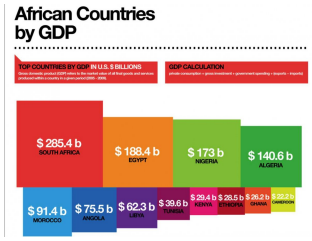
Billion dollar o-gram

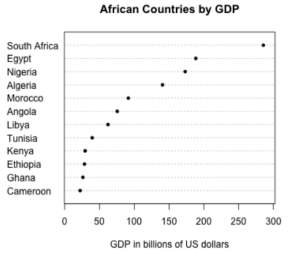


Word Clouds vs dotcharts

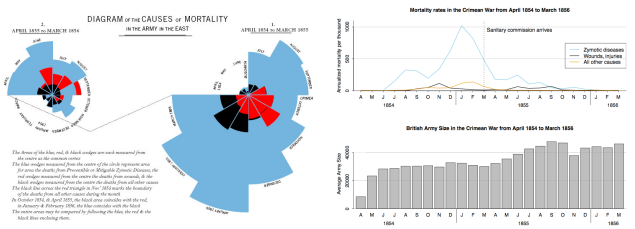


Rectangle Map vs Dot chart

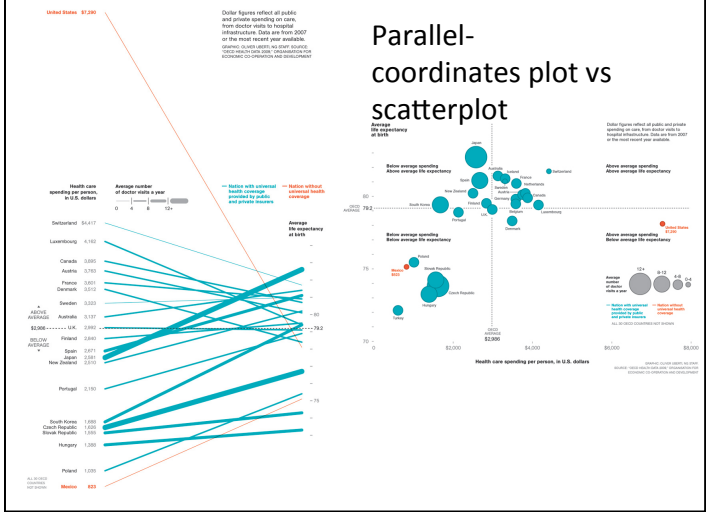




Windrose vs density plots and bar plots



Parallel-coordinates plot vs scatterplot



General Graphics References

- Cleveland,*The elements of graphing data*
- Tufte, *The visual display of quantitative information*
- Wainer, *Graphic discovery: A trout in the milk and other visual adventures*
- Wilkinson, *Grammar of Graphics*
- Yau, *Visualize This: The flowing data guide to design, visualization, and statistics*

R Graphics References

- Murrell, *R Graphics, 2nd edition*
- Sarkar, *Multivariate data visualization with R*
- Wickham, *ggplot2 – Elegant graphics for data analysis*