15.S60 visualization in R

Software Tools for Operations Research

January 14, 2014

why visualization?

"the picture-examining eye is the best finder we have of the wholly unanticipated"

-John Tukey

see relationships, structures, distributions, outliers, patterns, behaviors, dependencies, outcomes

high level outline

understand your data + intro to ggplot

about 90 minutes

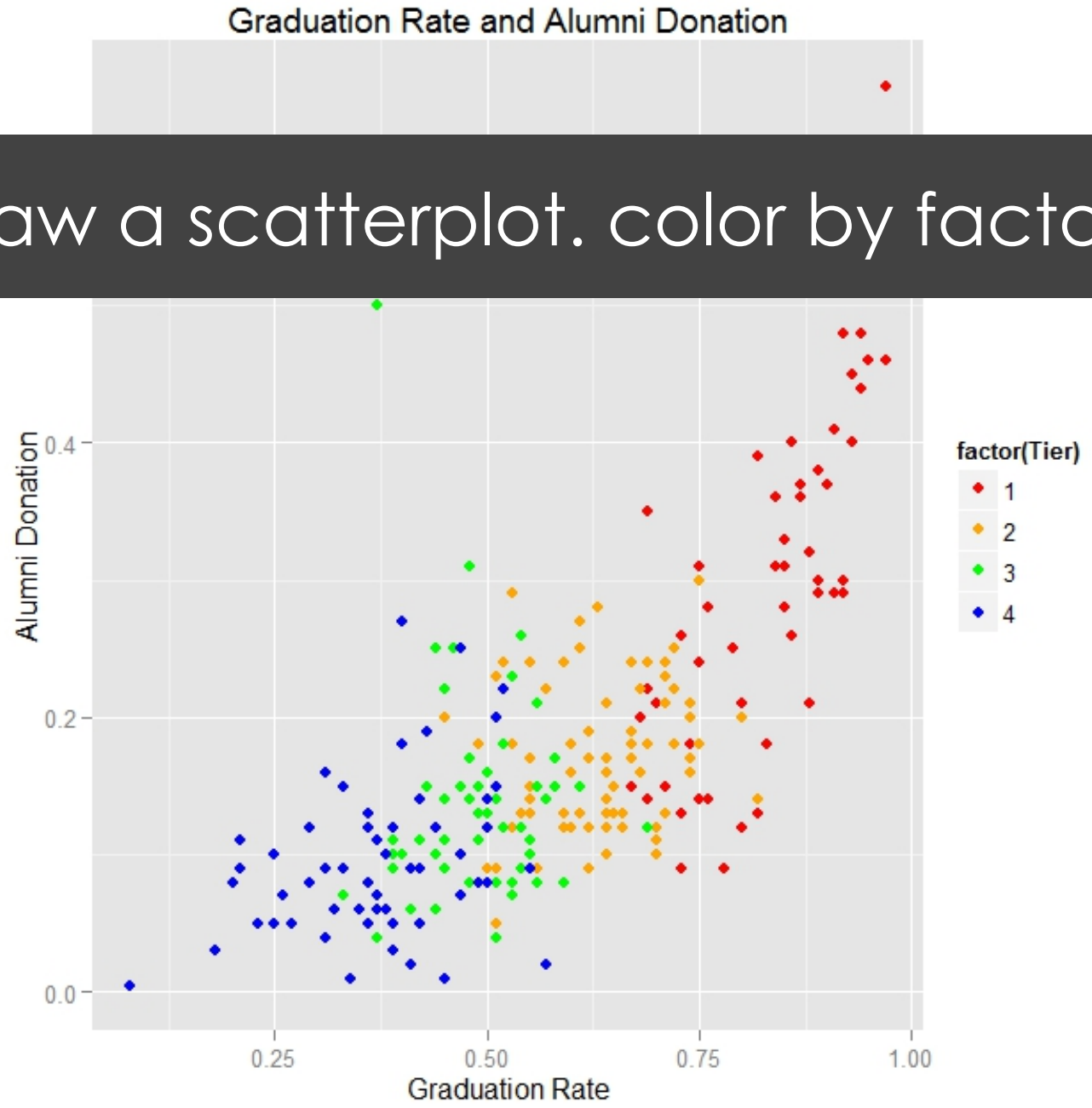
understand your model

about 40 minutes

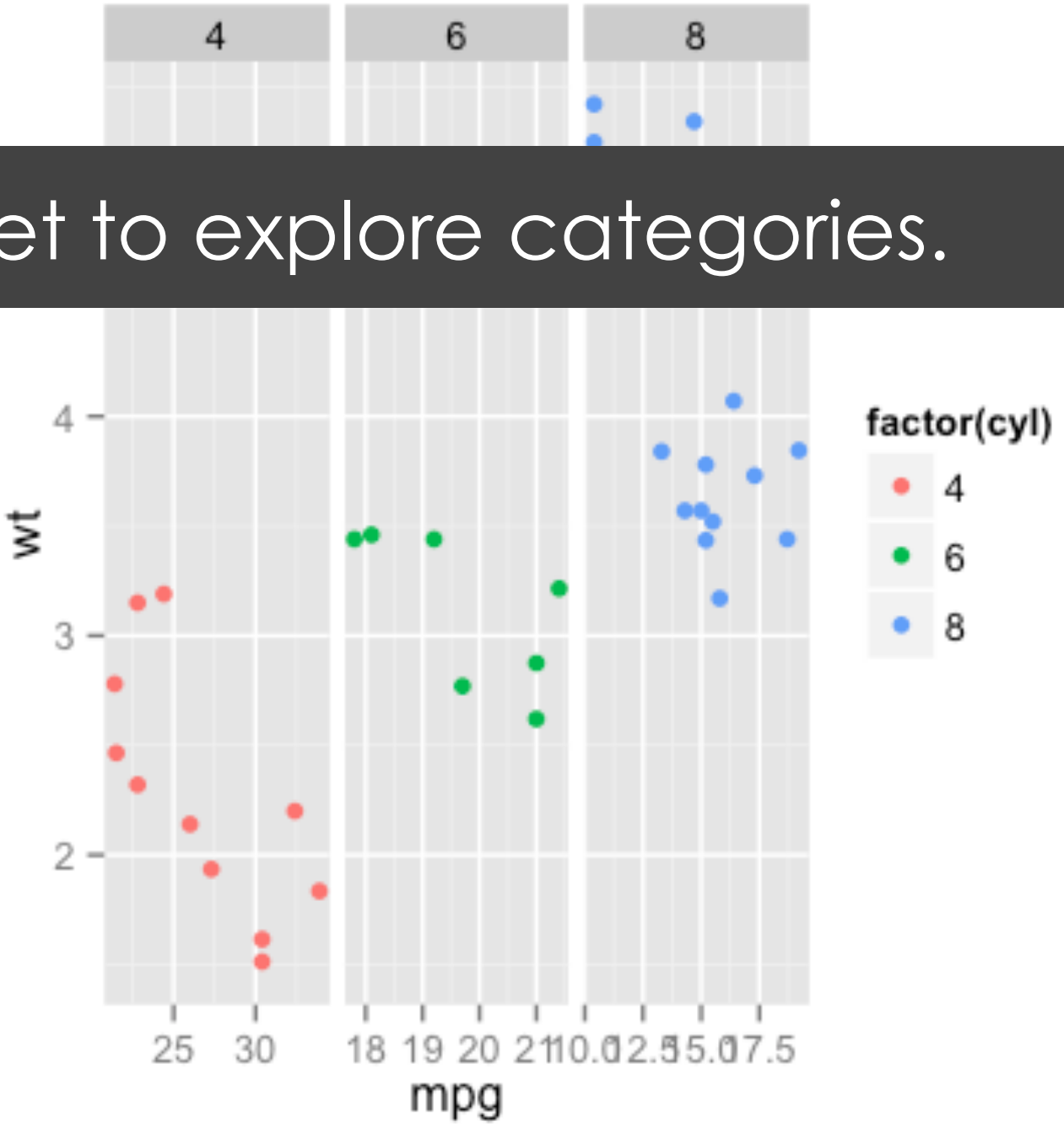
communicate your findings + best practices

About 20 minutes

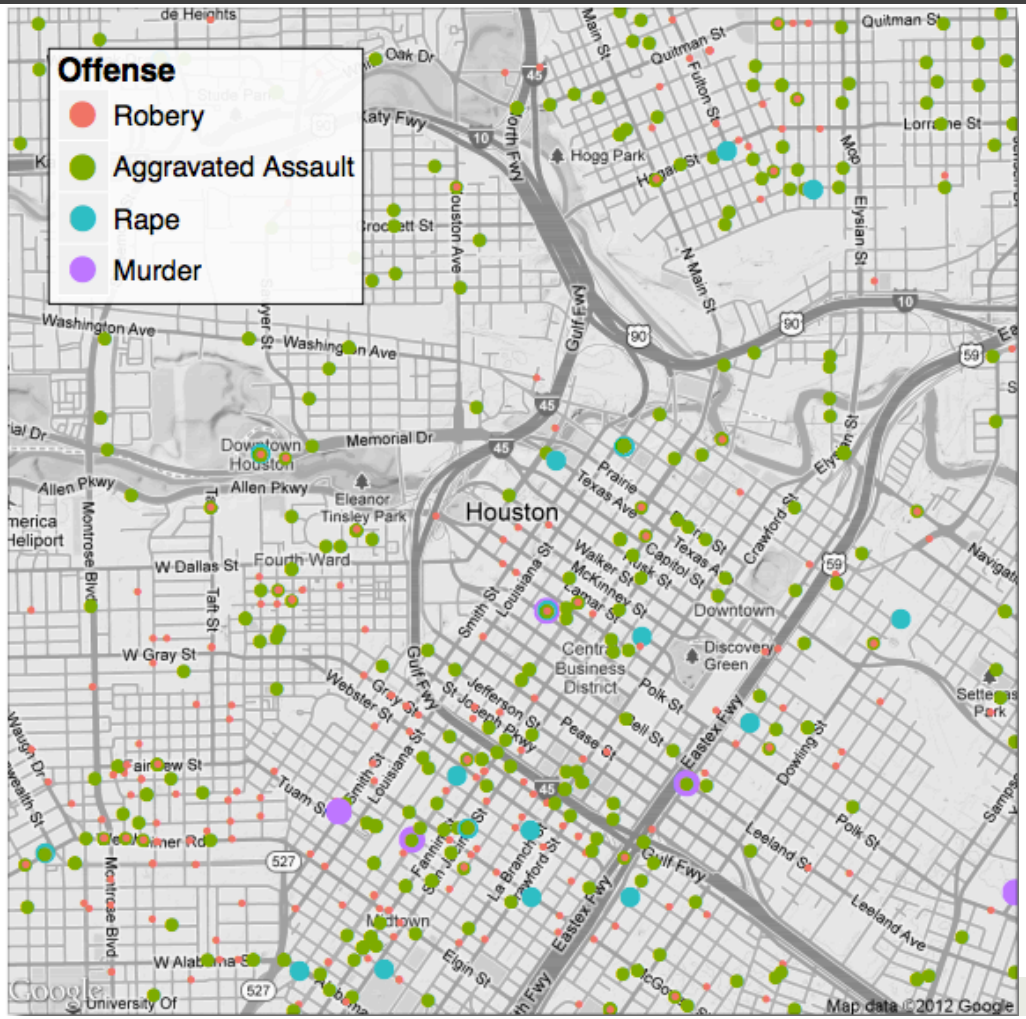
draw a scatterplot. color by factor.



facet to explore categories.

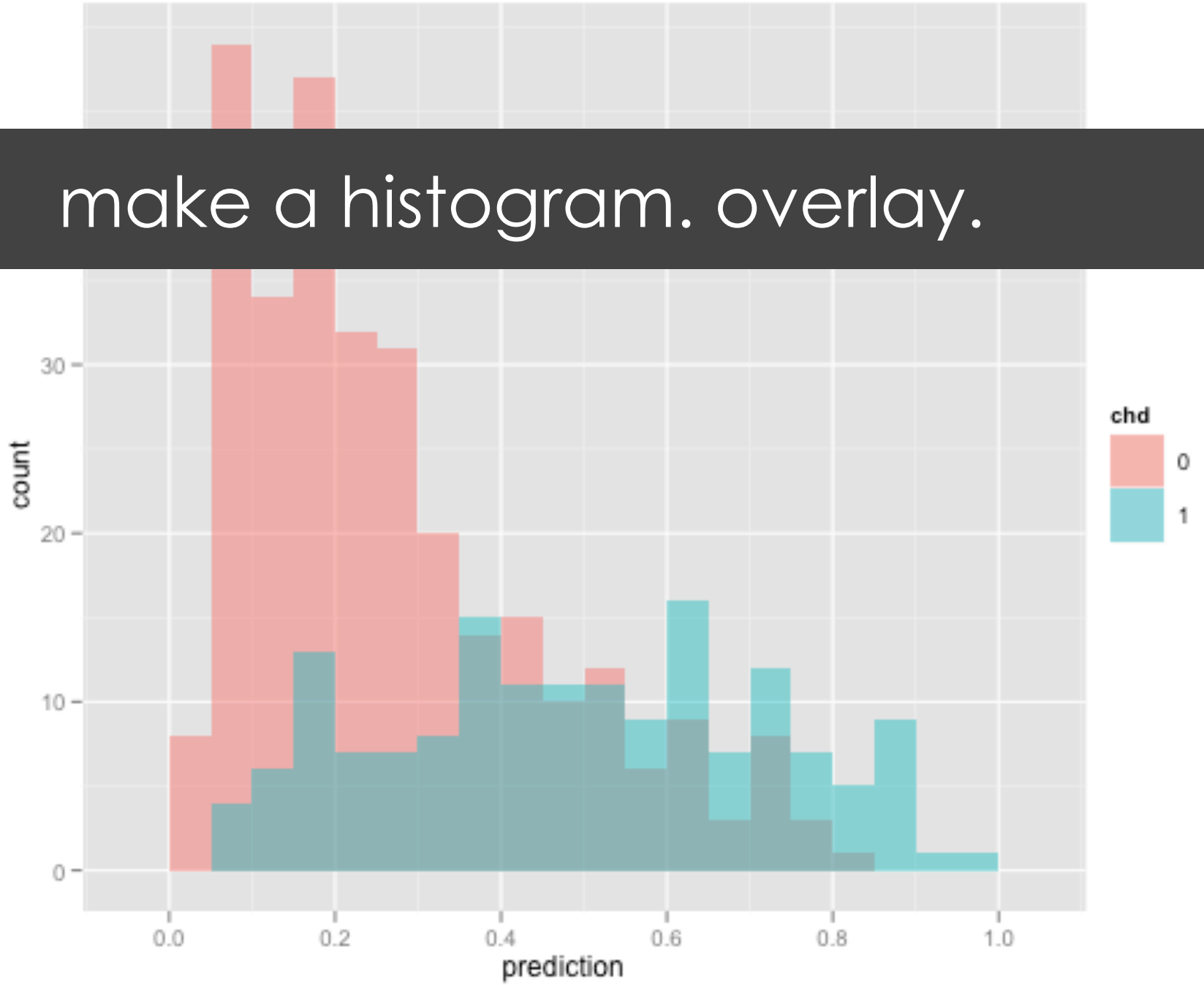


add data to a map. edit point size.

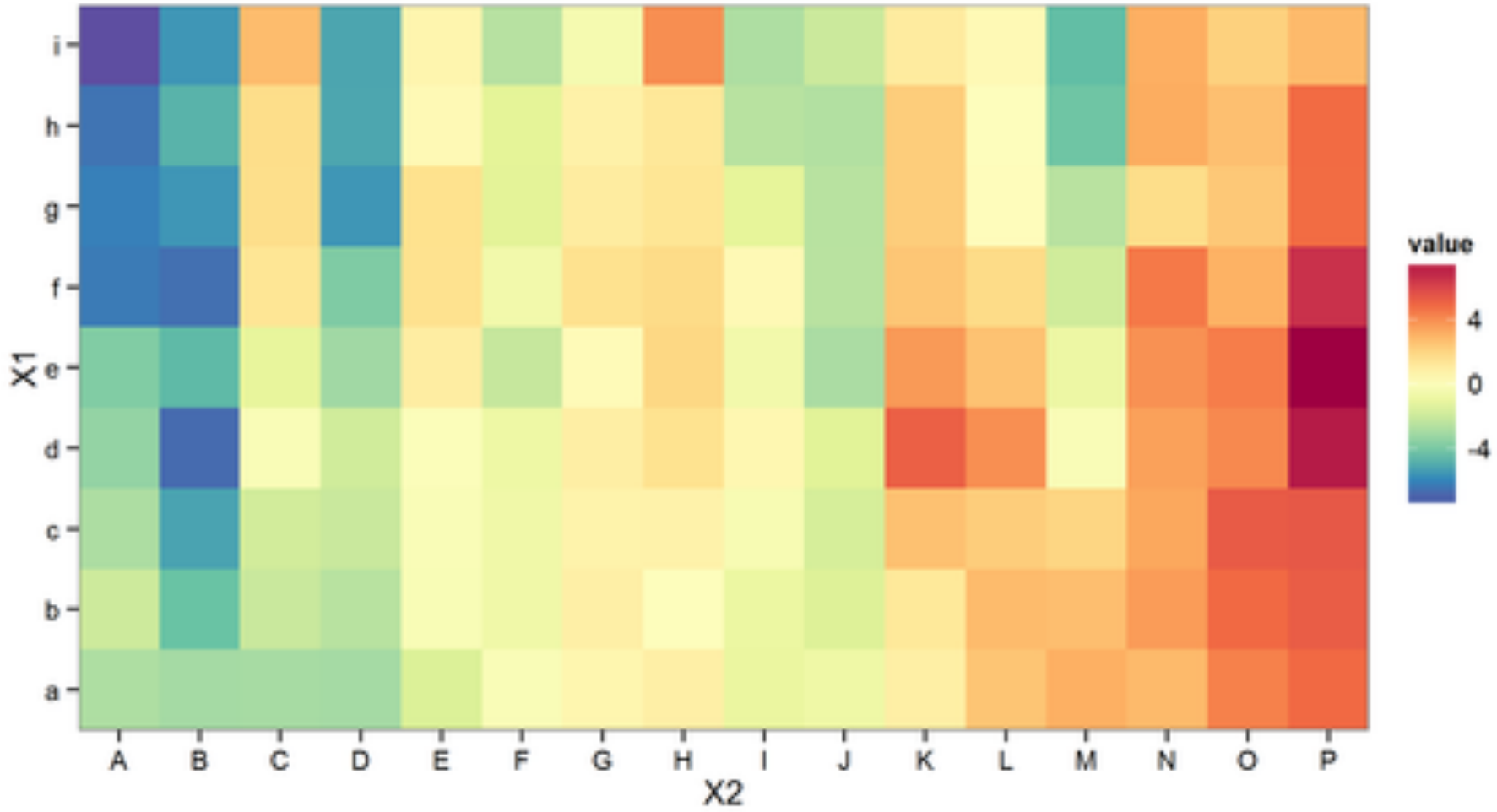


Crime in Houston TX

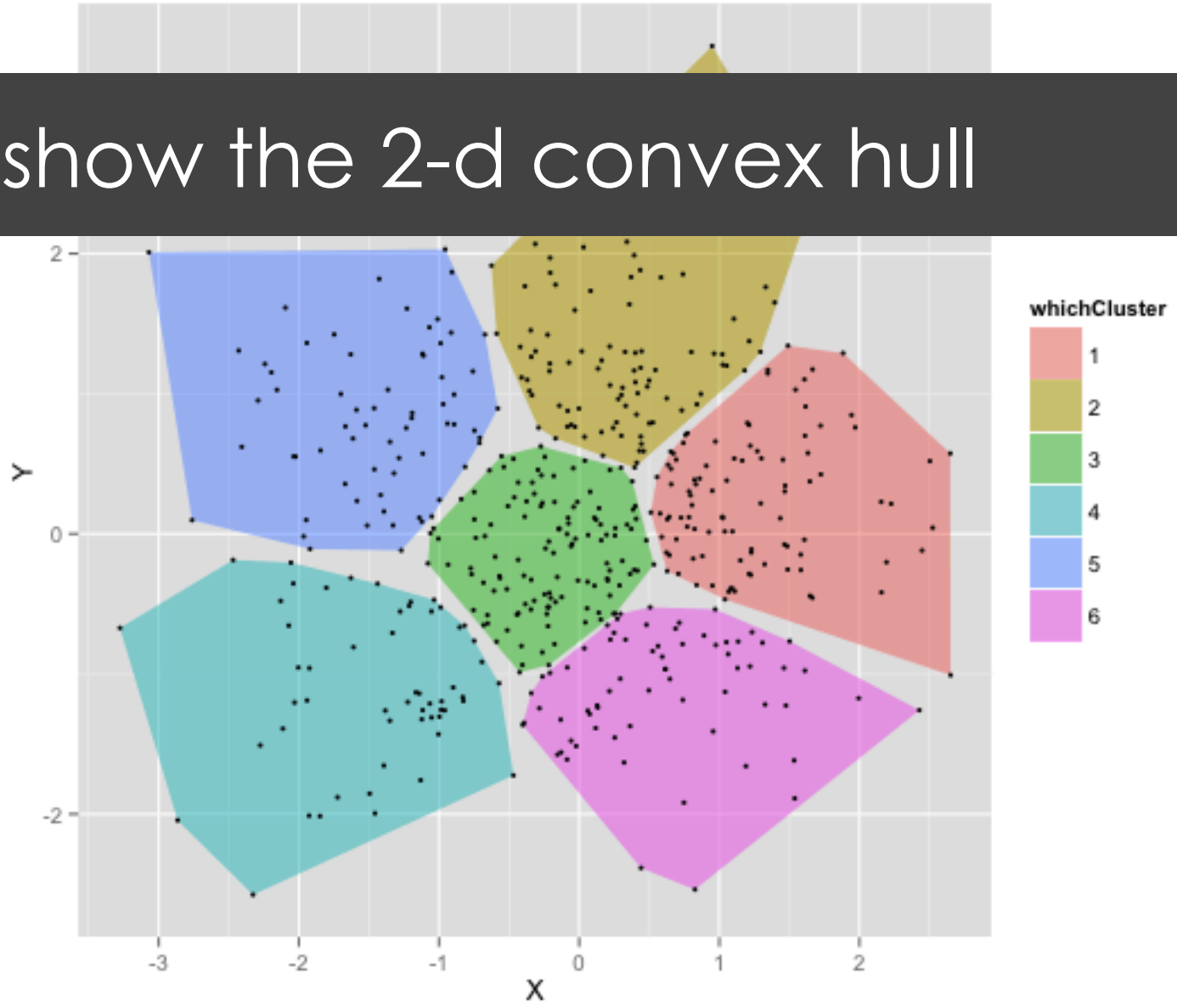
make a histogram. overlay.



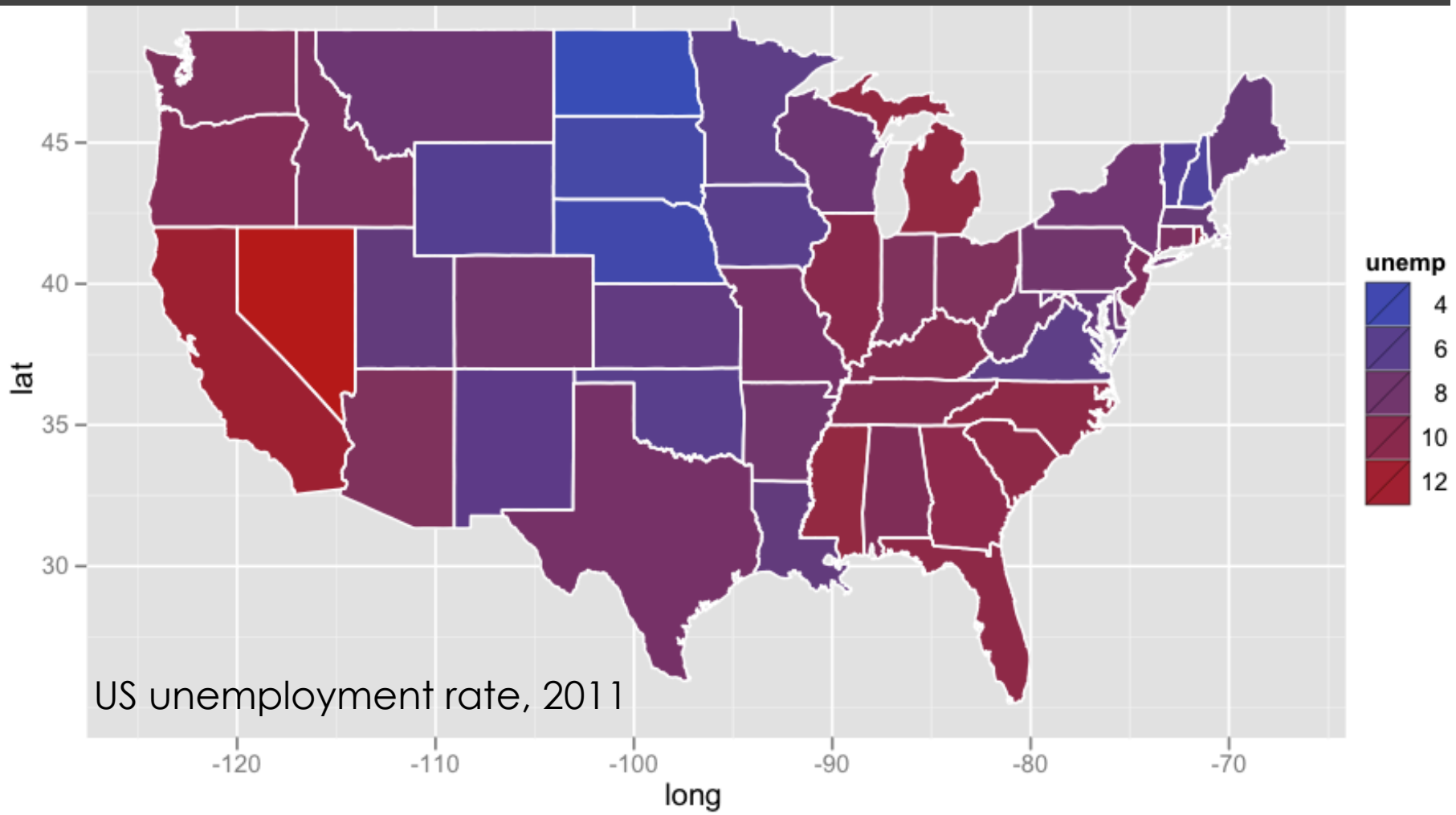
make a heat map. choose colors.



show the 2-d convex hull

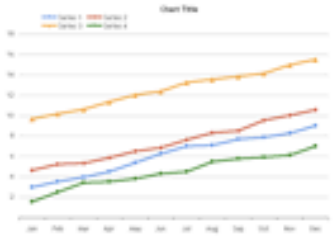


color a map according to data

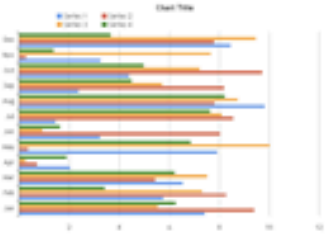


choose the right visualization

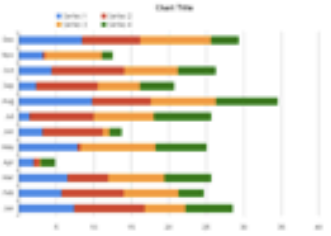
Line chart



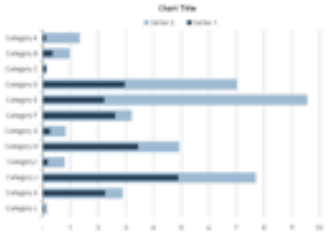
Bar chart



Stacked bar chart

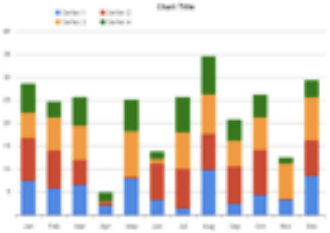


Bullet bar chart



Column chart

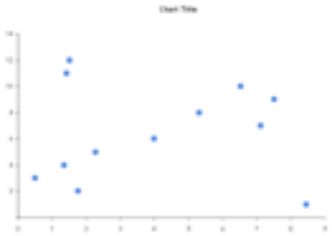
Stacked column chart



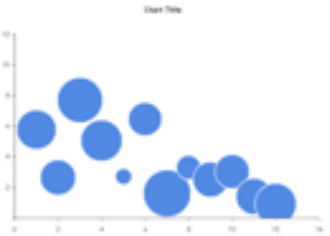
Pie chart

Pie chart with highlight

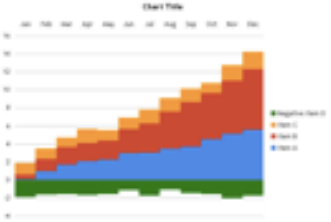
Scatterplot chart



Bubble chart



Stacked column volume chc



Stacked column volume wit

datasets we'll explore today

anscombe's quartet

iris dataset

election polling data

hubway data

hubwaydatachallenge.org

- do Hubway cyclists only ride downhill?
- are all Hubway rentals after 2 am only by people under 25?
- how do weekday and weekend patterns differ?

finish class with our own version of hubway visualization challenge!

section 1 – understand data

- scatterplots
- clusters
- map data
- histograms
- heat maps

Others you could investigate: box plots, cluster dendrograms, time series data, networks,…

what is ggplot?

"ggplot2 is a plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts. It takes care of many of the fiddly details that make plotting a hassle (like drawing legends) as well as providing a powerful model of graphics that makes it easy to produce complex multi-layered graphics."

-Hadley Wickham, creator, www.ggplot2.org

what is a data visualization?

A mapping of data properties to visual properties

Data properties are usually numerical or categorical

Visual properties can be (x,y) coordinates, colors, sizes, shapes, heights, etc

example: motor trends

mtcars:

mpg ➔ y-axis

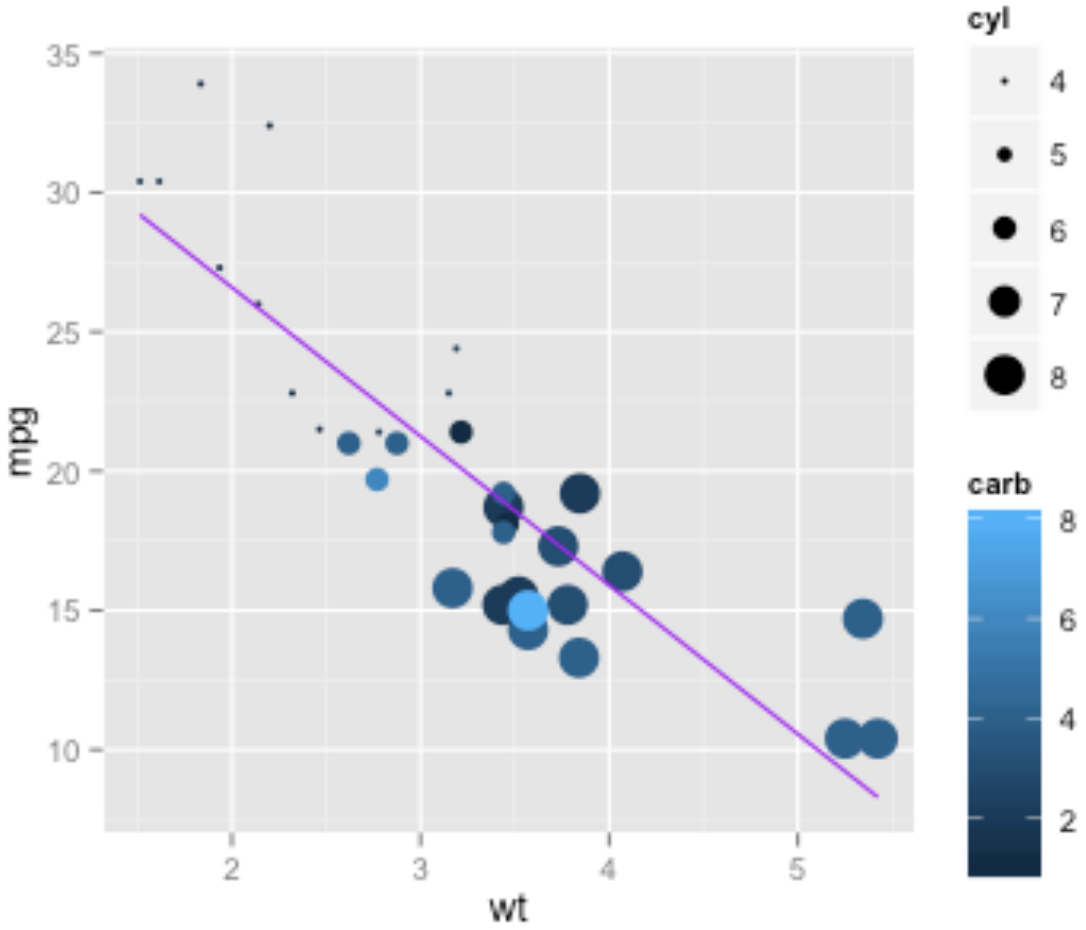
wt ➔ x-axis

carb ➔ color

cyl ➔ size

Regression equation

➔ purple line



graphics in base R vs ggplot

In base R, each mapping of data properties to visual properties is its own special case

- Graphics composed of simple elements like points, lines
- Difficult to add elements to existing plots

In ggplot, the mapping of data properties to visual properties is done by adding layers to the plot, starting with the raw data

grammar of graphics

ggplot graphics consist of at least 3 elements:

1. Data, in a data frame

2. Aesthetic mapping (aes) describing how variables in the data frame are mapped to graphical attributes

   color, shape, scale, x-y axes, subset groupings…

3. Geometric objects (geoms) determine how values are rendered graphically

   points, lines, boxplots, bars, polygons, …

example: motor trends

mtcars:

mpg ➔ y-axis

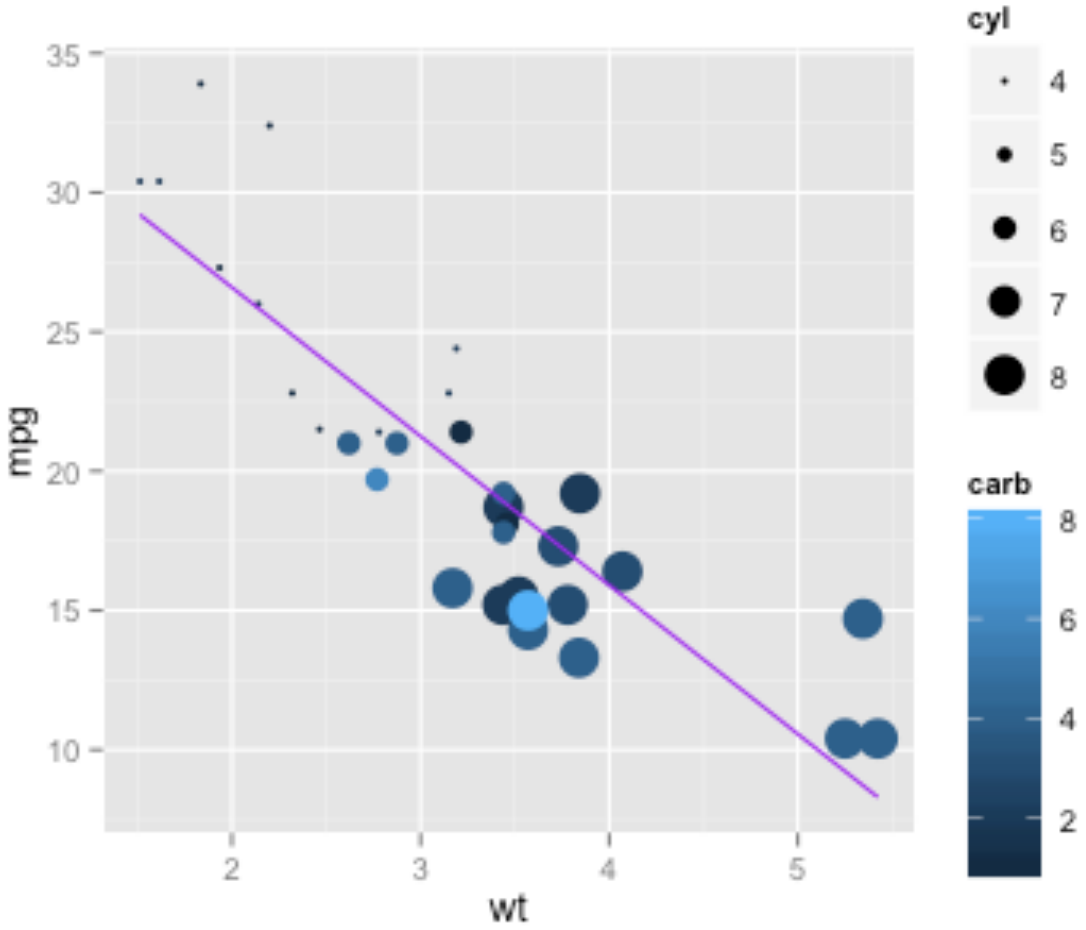
wt ➔ x-axis

carb ➔ color

cyl ➔ size

Regression equation

➔ purple line



other potential layers

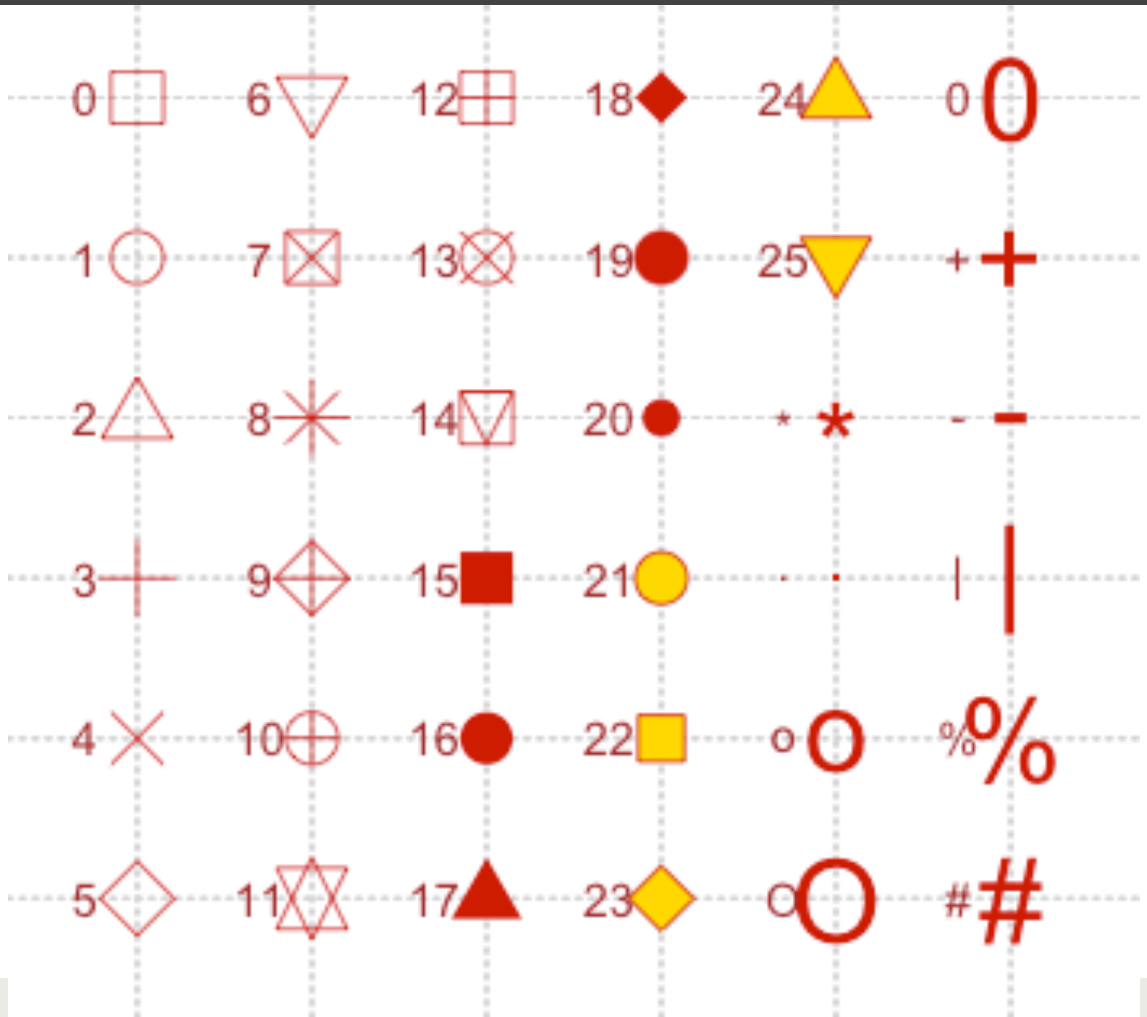
1. Statistical Transformations (stat)
   (smoothing, binning in a histogram, regression lines,...)

2. Facets to create lattices of plots
   (ex: plot each patient individually)

3. A coordinate system (cartesian coordinates, log coordinates, polar coordinates, map projections, …)

4. Scales (adjust size scales, color scales, etc)

point shapes in R



exercise 1

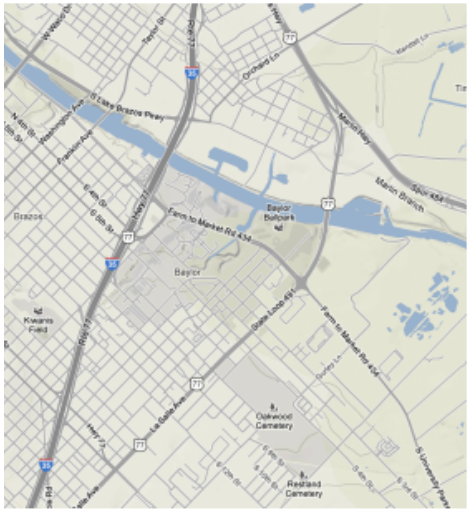
Plot sepal length vs. petal length in a scatterplot and title the graph.

exercise 2

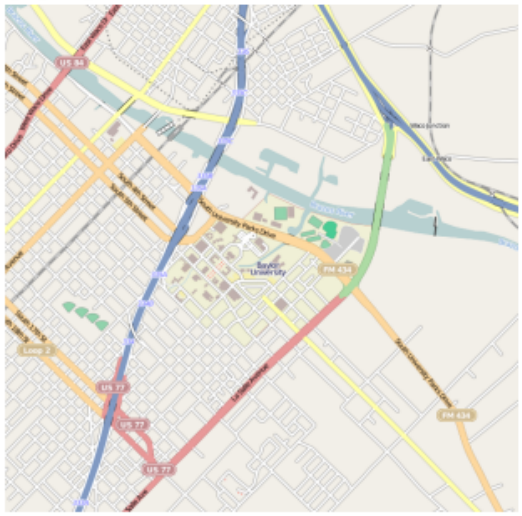
Change the size of points, by petal width.

the ggmap package

source = “google”,
maptype = “terrain”



source = “osm”



source = “stamen”,
maptype = “toner”

source = “stamen”
maptype =
“watercolor”

exercise 3

Plot the Hubway stations on a map of Boston.

Change the station size to be relative to the total number of trips out of the station.

Hint: First create a frequency table of trips by start station using “table” and turn this into a data frame. Then merge frequency and station data frames.

exercise 4

Plot a histogram of trip length.

Make bin sizes 10 minutes long.

Facet it by subscription type.

What do you notice?

section 2 – understand your model

- add a regression line to a scatterplot
- visualize output of clustering model by drawing convex hulls of clusters
- plot the output of logistic regression on a map

Other model output to visualize: CART trees and partitions, Markov processes, linear regression diagnostic plots, branch & bound trees, plots of IP solver time to optimality, …

exercise 5

Draw a 99% confidence interval instead.

Make the regression line hot pink and large enough to show up.

the electoral college

The United States has 50 states

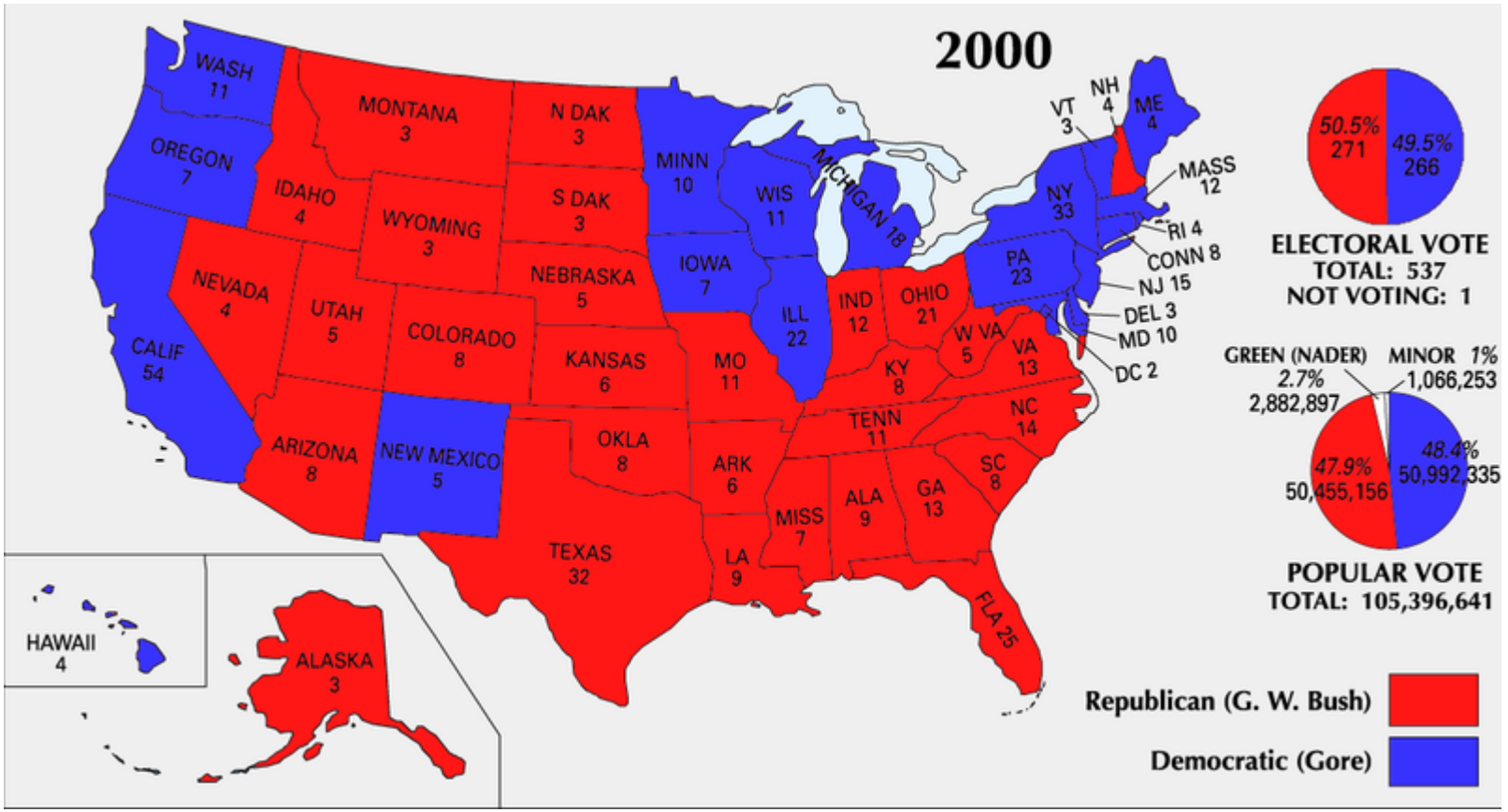
Each assigned a number of *electoral votes* based on population

- Most votes: 55 (California)
- Least votes: 3 (multiple states)
- Reassigned periodically based on population change

Winner takes all: candidate with the most votes in a state gets all its electoral votes

Candidate with most electoral votes wins election

2000 Election: Bush vs. Gore



election prediction

Goal: Use polling data to predict state winners

Then-*New York Times* columnist Nate Silver famously took on this task for the 2012 election

polling data

Dependent variable: **Republican**

1 if republican won state, 0 if democrat

SurveyUSA: polled R% - polled D%

DiffCount: polls with R winner - polls with D winner

exercise 6

Plot prediction results as a gradient according to logistic regression probabilities. Change the legend's title to "Prediction 2012" using the name parameter.

How do you think visualizations should best represent uncertainty? Discuss with your neighbor.

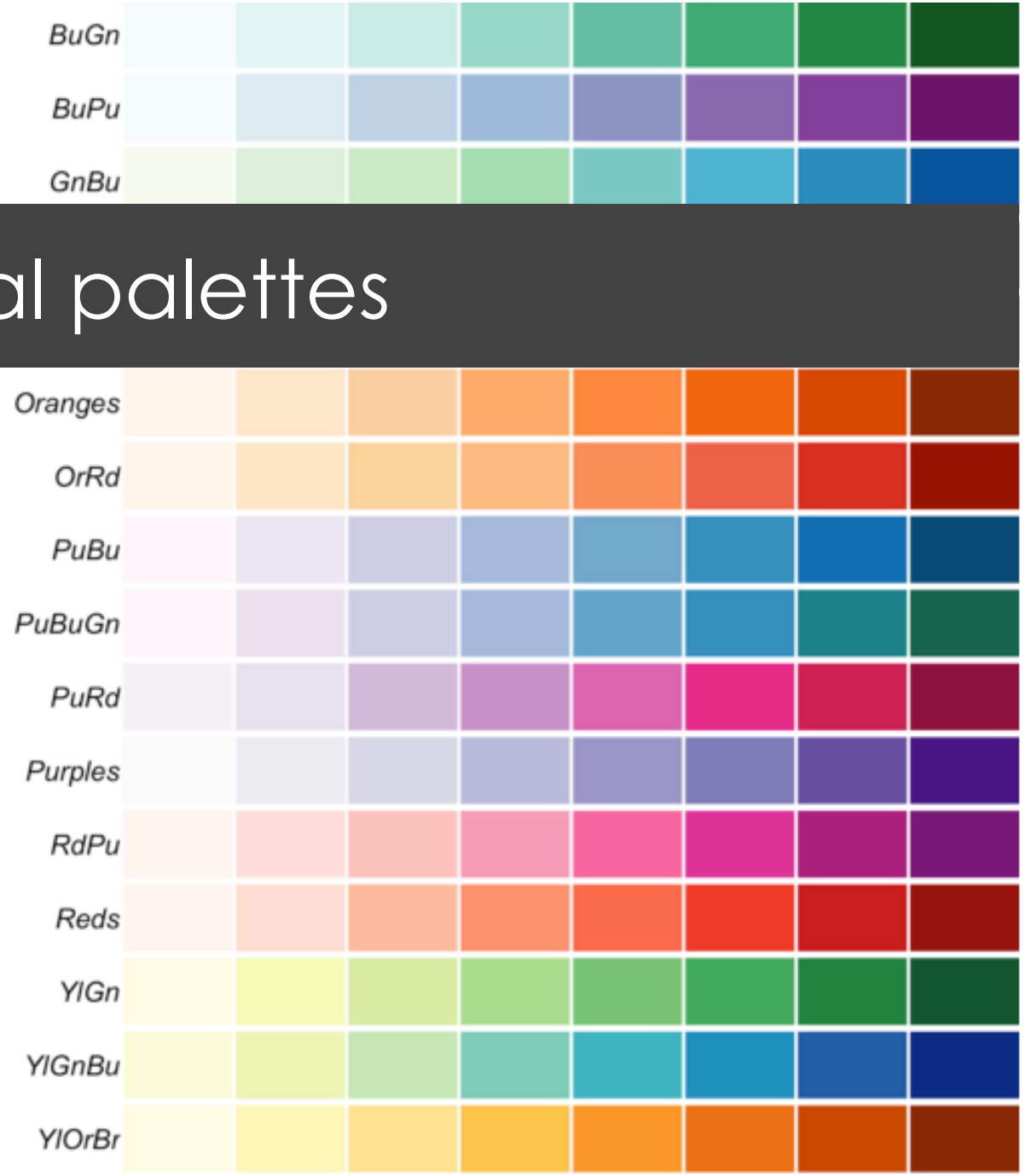
section 3 – communicate

- colors
- best practices
- what not to do
- next steps: making visualizations dynamic, interactive

Others you could investigate: fonts, themes, text annotations, …

sequential palettes

For ordered data that progresses from low to high values

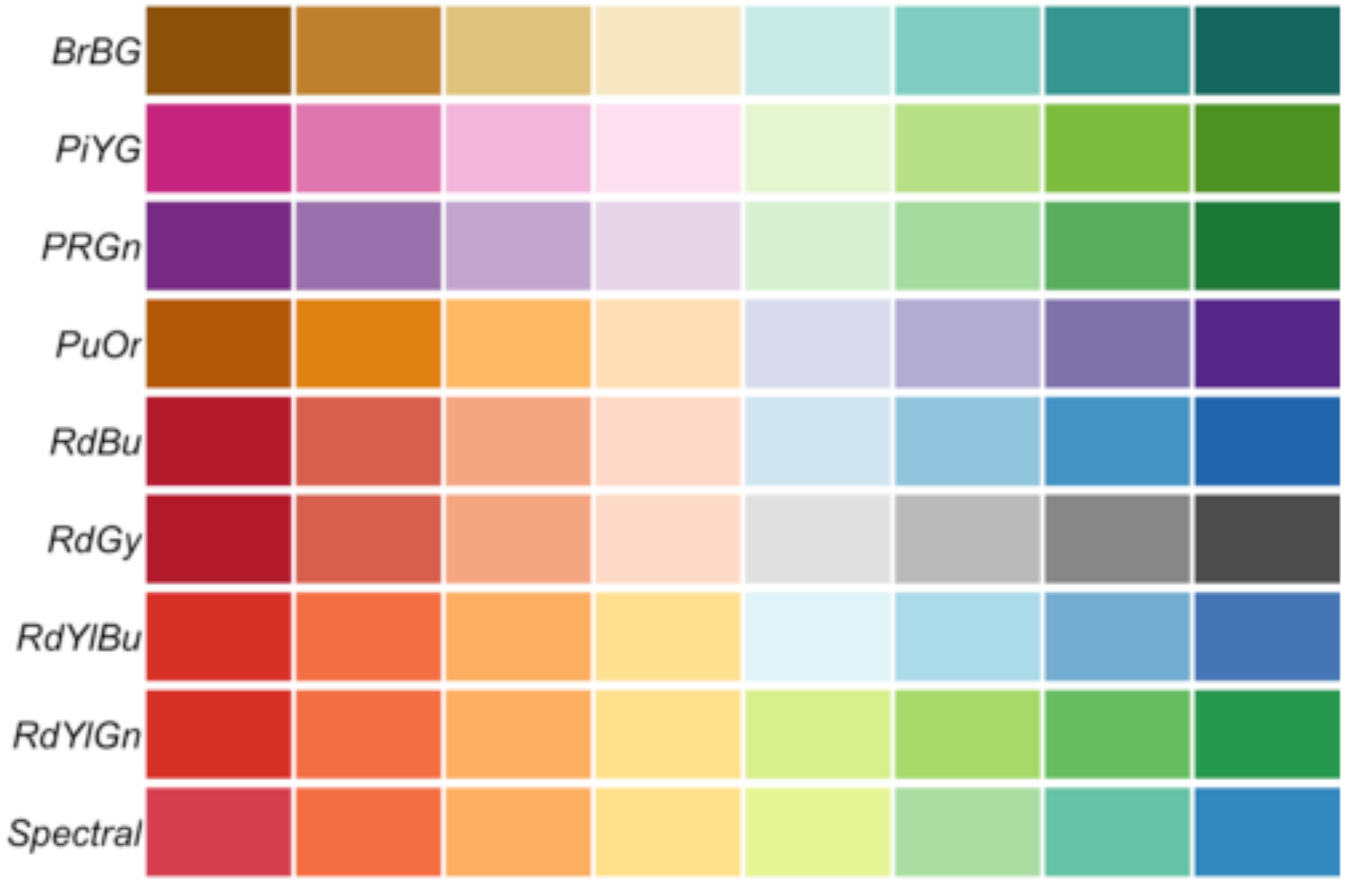


qualitative palettes

Best for categorical data

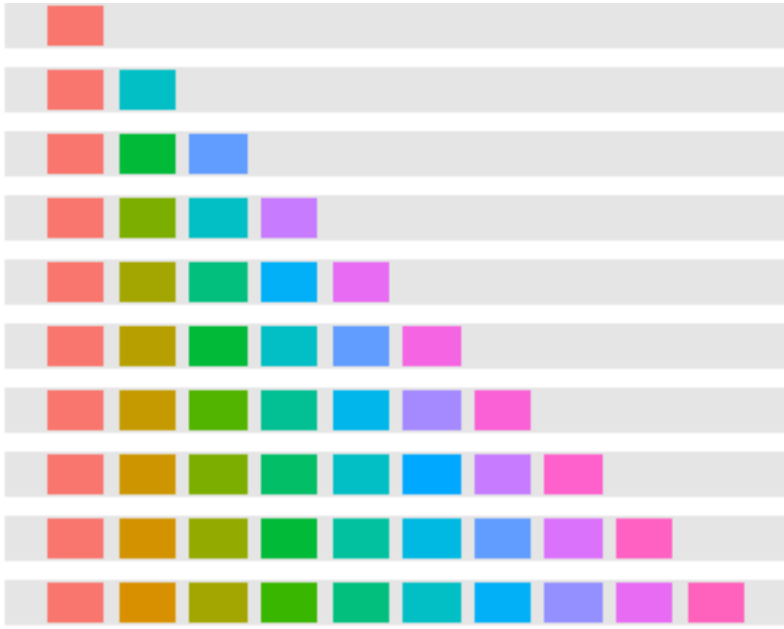
diverging palettes

Emphasize extremes



ggplot default color selection

Maximal dispersion along ggplot's color wheel



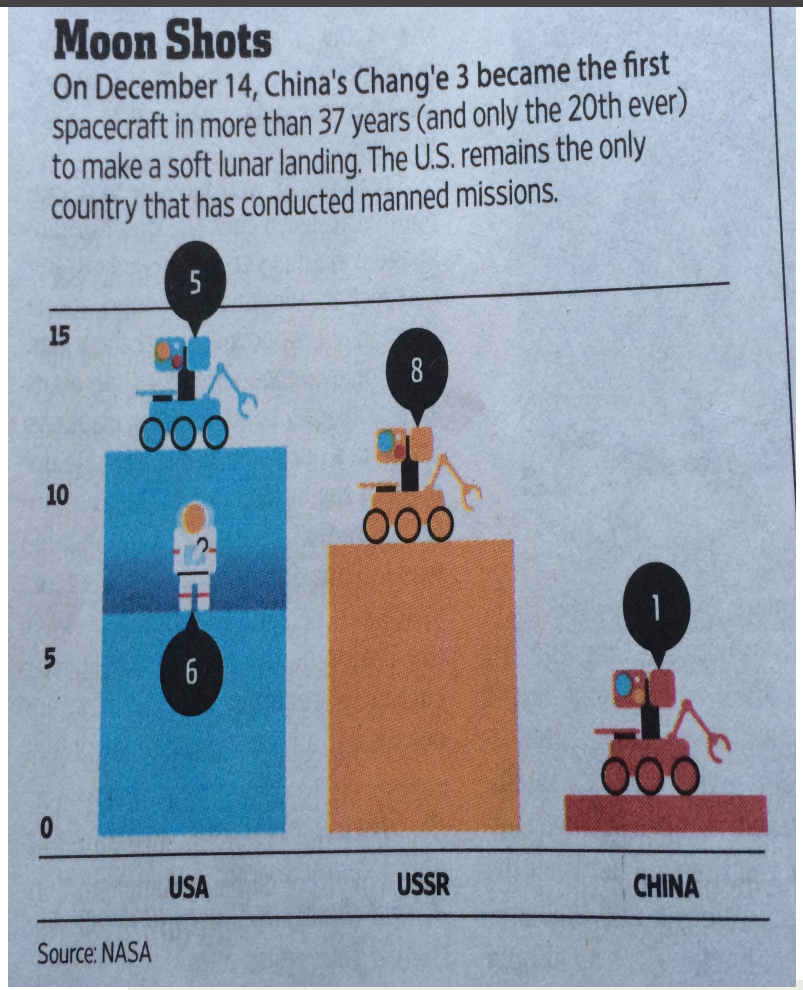
a good visualization…

- accurately conveys data
- distills and summarizes
- helps to discern relationships
- takes into account audience considerations

bad visualizations…

- too much information
- display choices distort reality
- inconsistent scales, ordering, placement
- too many colors/patterns/styles
- absolute values vs. percentages

bad visualization



Height of bar is total missions to moon, which you can try to read off the y-axis or manually add numbers - redundant

Two types of missions stacked in one bar makes it hard to compare columns

Overall - too complicated

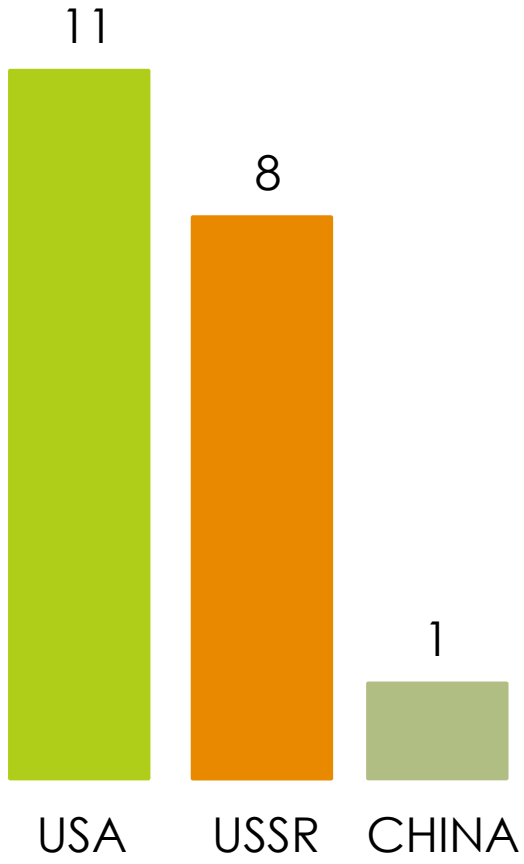
better version

Raw data is simple:

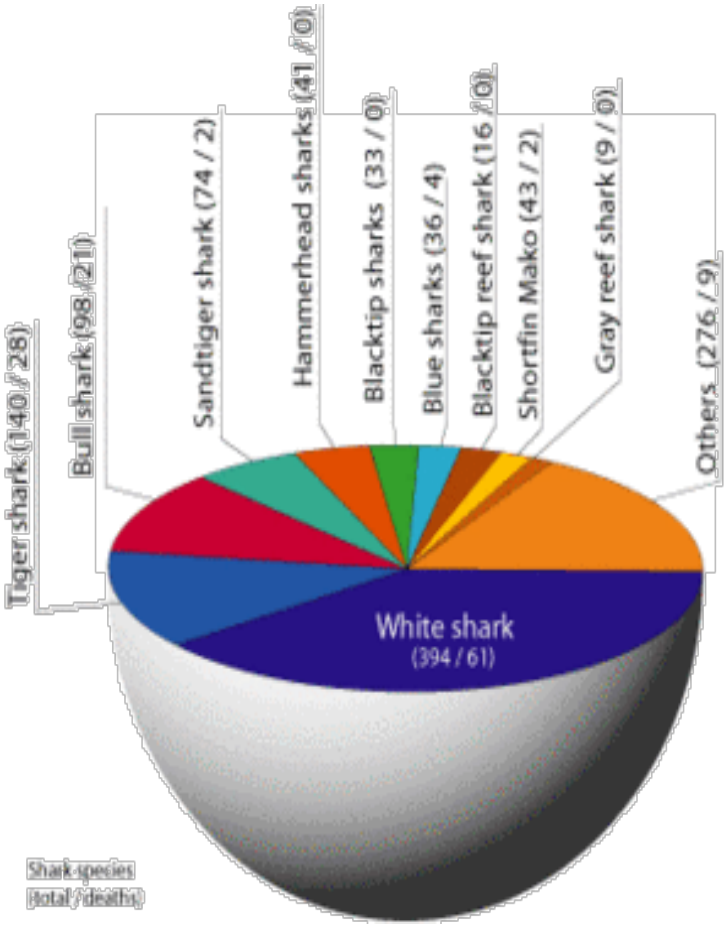
```
USA (5,6), USSR (8,0),
CHINA (1,0)
```

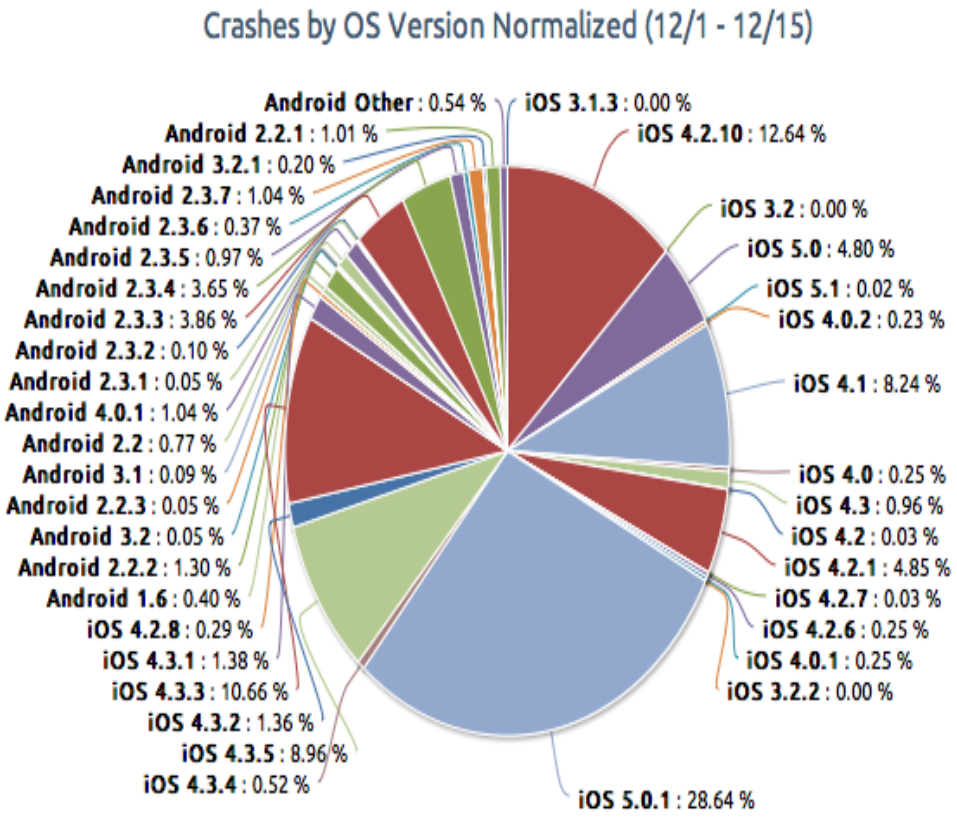
Keep it simple - plot total missions only, unambiguous

Numbers provide details, relative scale easy to determine at a glance



pie charts

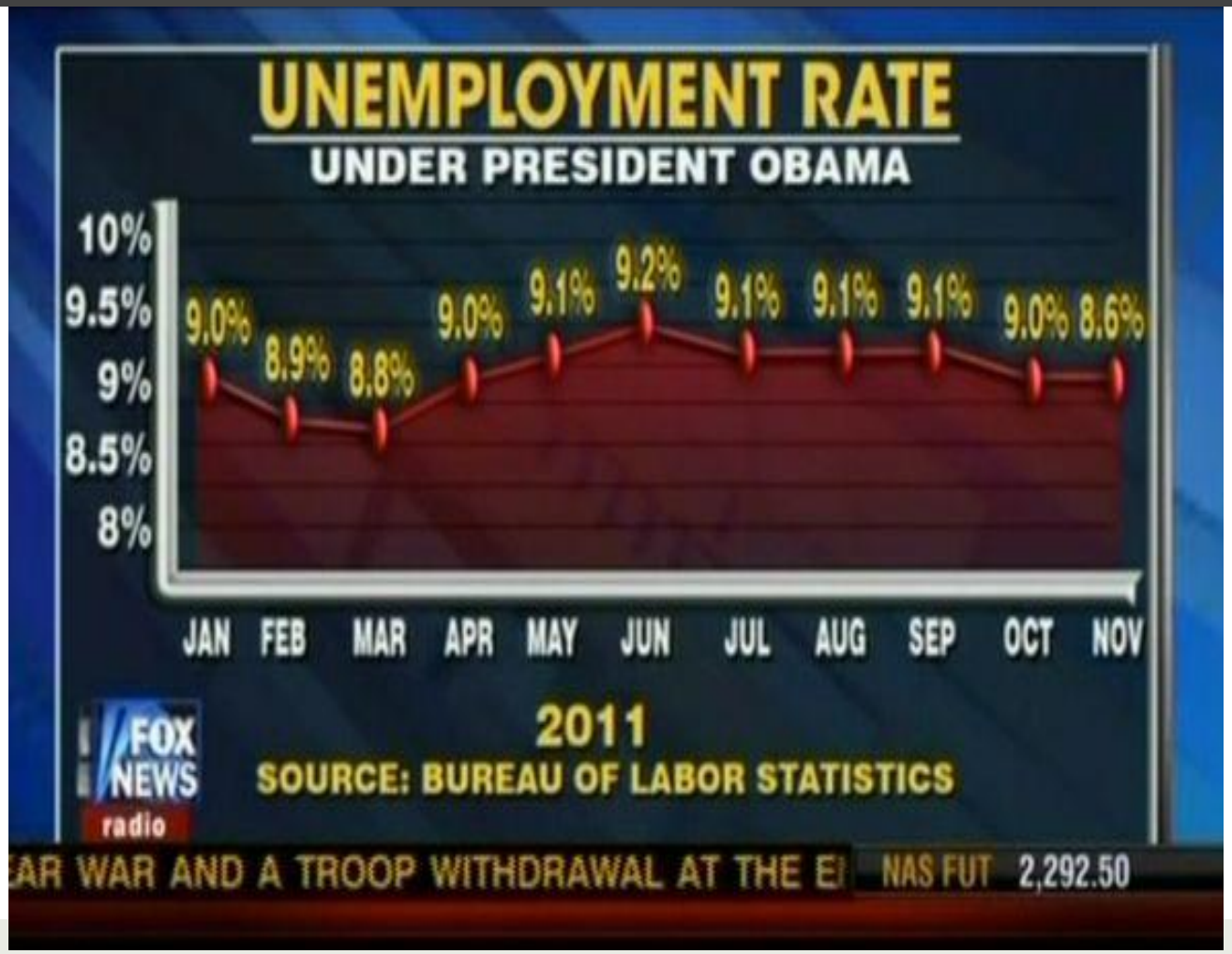

Tiger shark (140 / 28)
Bull shark (98 / 21)
Sandtiger shark (74 / 2)
Hammerhead sharks (41 / 0)
Blacktip sharks (33 / 0)
Blue sharks (36 / 4)
Blacktip reef shark (16 / 0)
Shortfin Mako (43 / 2)
Gray reef shark (9 / 0)
Others (276 / 9)
White shark
(394 / 61)
Shark species
Total / deaths


Crashes by OS Version Normalized (12/1 - 12/15)
Android Other : 0.54 %
Android 2.2.1 : 1.01 %
Android 3.2.1 : 0.20 %
Android 2.3.7 : 1.04 %
Android 2.3.6 : 0.37 %
Android 2.3.5 : 0.97 %
Android 2.3.4 : 3.65 %
Android 2.3.3 : 3.86 %
Android 2.3.2 : 0.10 %
Android 2.3.1 : 0.05 %
Android 4.0.1 : 1.04 %
Android 2.2 : 0.77 %
Android 3.1 : 0.09 %
Android 2.2.3 : 0.05 %
Android 3.2 : 0.05 %
Android 2.2.2 : 1.30 %
Android 1.6 : 0.40 %
iOS 4.2.8 : 0.29 %
iOS 4.3.1 : 1.38 %
iOS 4.3.3 : 10.66 %
iOS 4.3.2 : 1.36 %
iOS 4.3.5 : 8.96 %
iOS 4.3.4 : 0.52 %
iOS 3.1.3 : 0.00 %
iOS 4.2.10 : 12.64 %
iOS 3.2 : 0.00 %
iOS 5.0 : 4.80 %
iOS 5.1 : 0.02 %
iOS 4.0.2 : 0.23 %
iOS 4.1 : 8.24 %
iOS 4.0 : 0.25 %
iOS 4.3 : 0.96 %
iOS 4.2 : 0.03 %
iOS 4.2.1 : 4.85 %
iOS 4.2.7 : 0.03 %
iOS 4.2.6 : 0.25 %
iOS 4.0.1 : 0.25 %
iOS 3.2.2 : 0.00 %
iOS 5.0.1 : 28.64 %

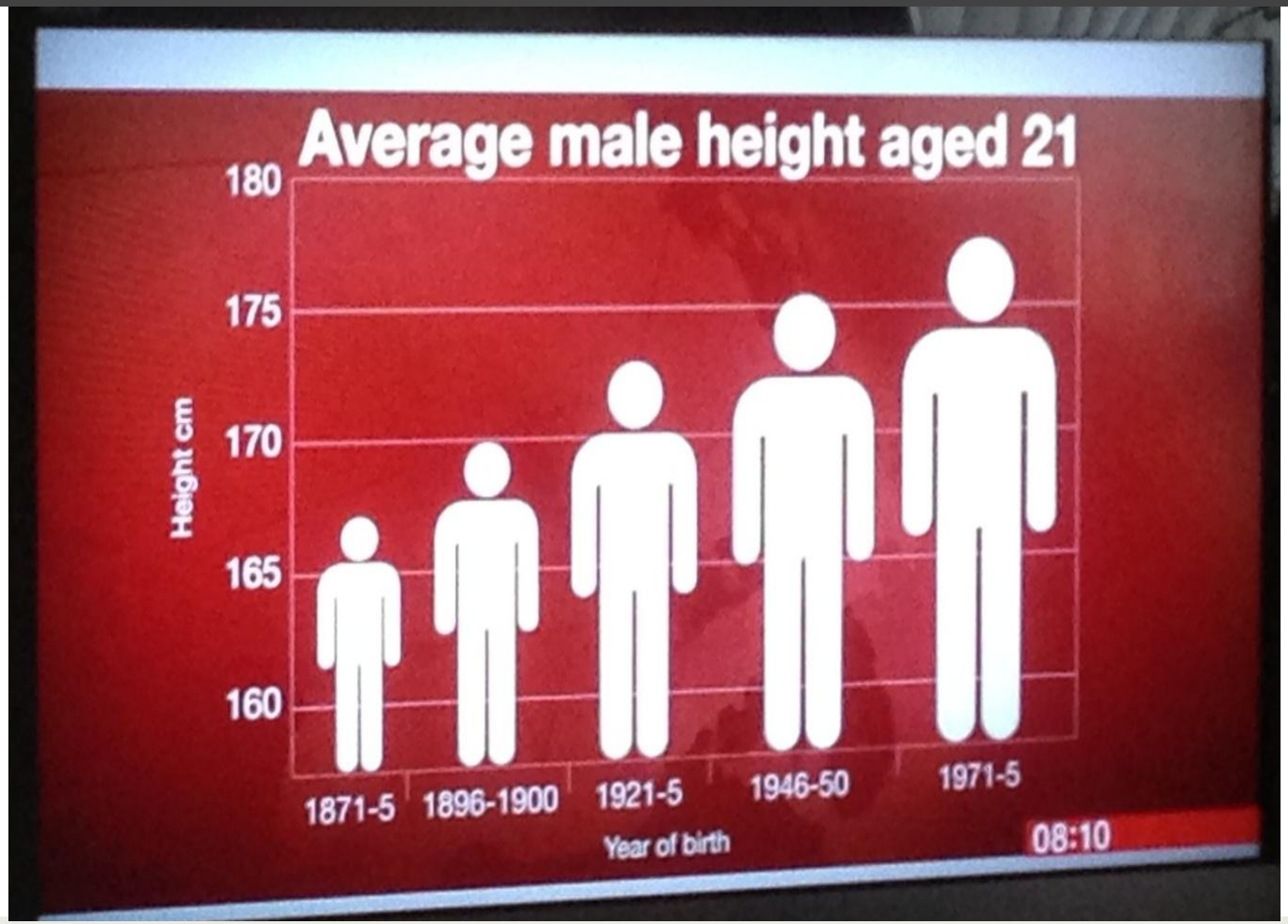
pie chart issues

- Inconsistent labeling
- Not all points can be labeled – data lost
- Colors meaningless
- Arc length meaningful, but hard to compare and eye is drawn to area
- 3D even worse: adds no information

misleading scales



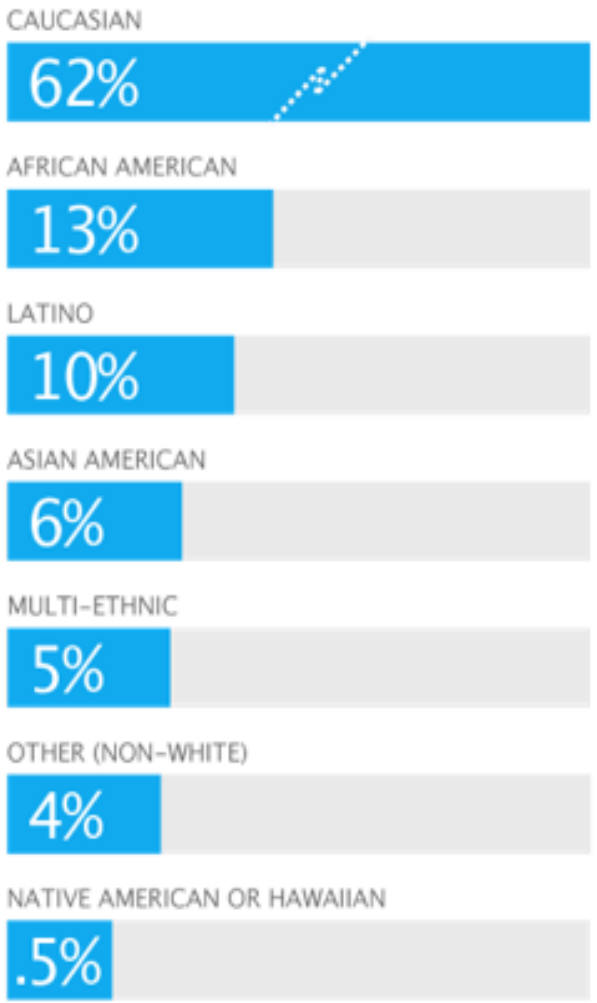
midgets to giants? not really…



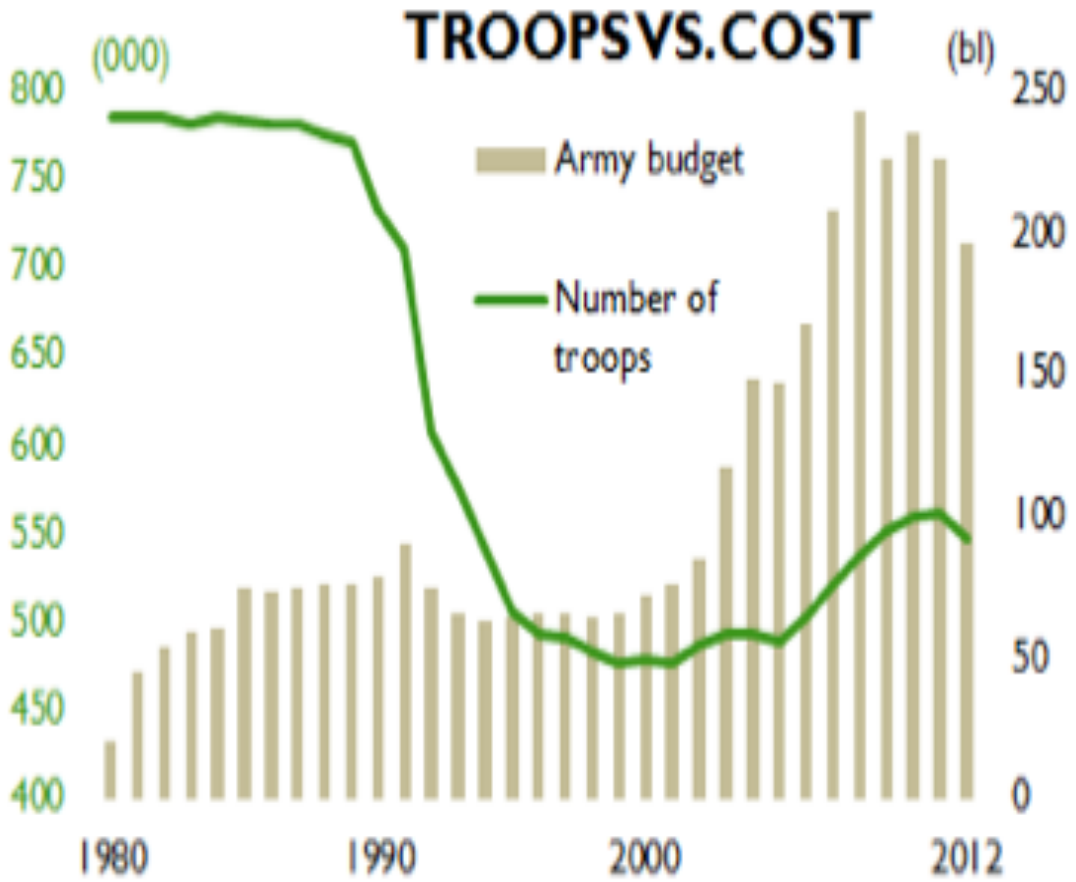
just plain ridiculous

Diversity for 2012 Corps Members

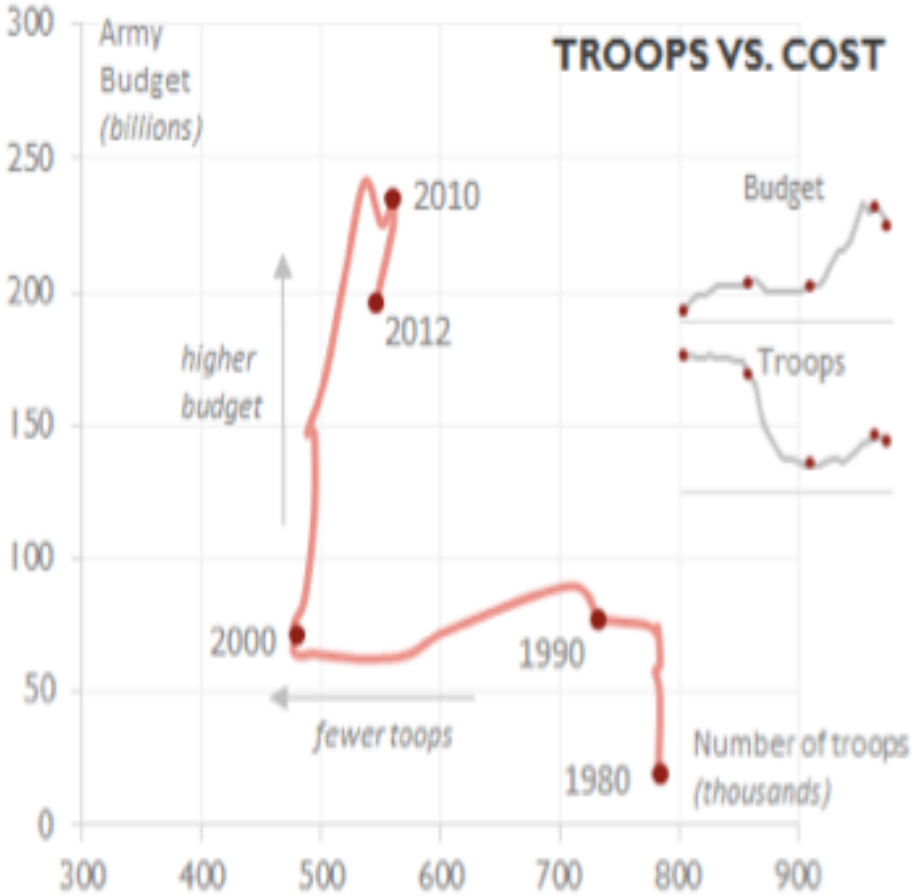
Total people of color: 38%



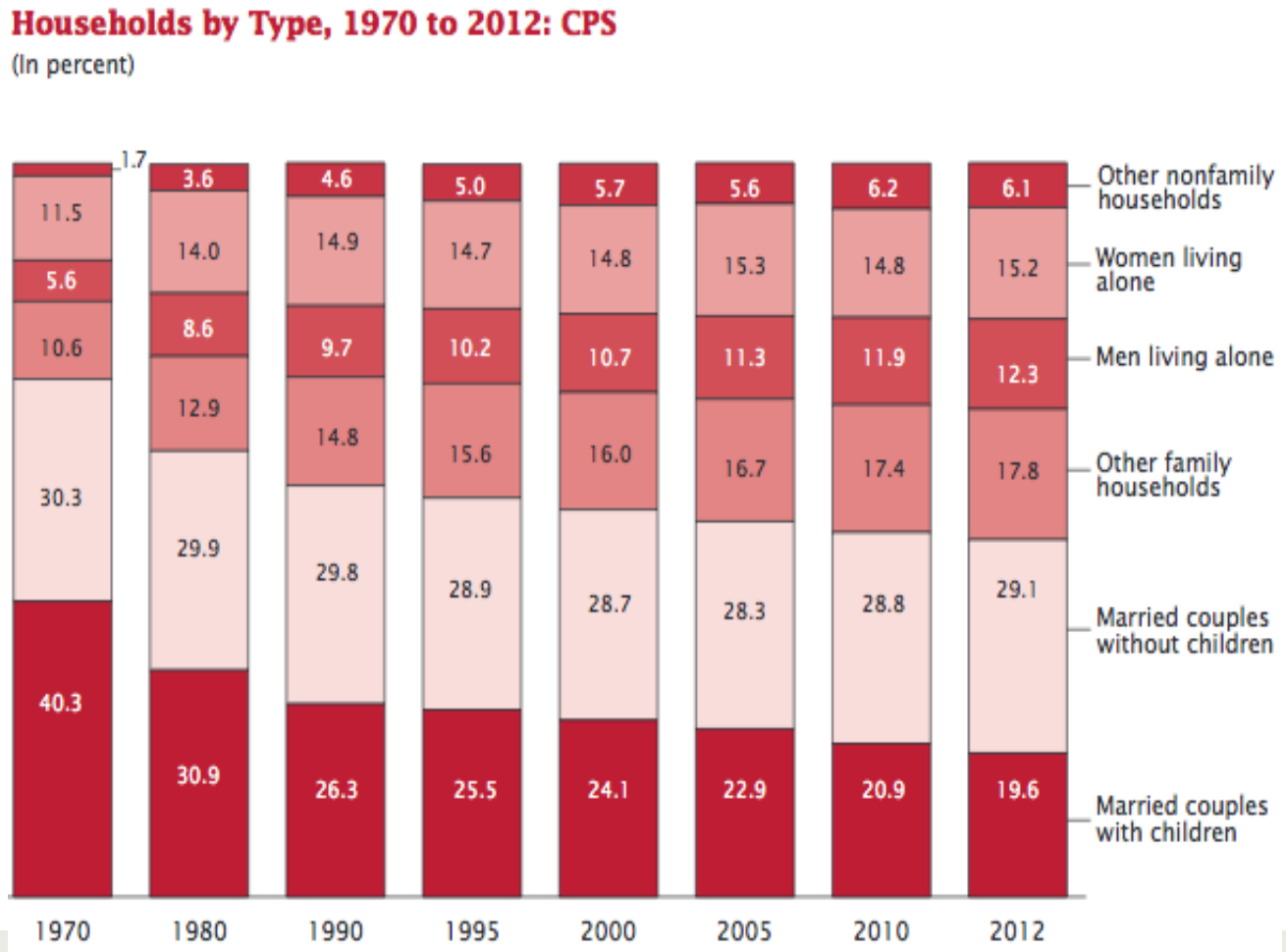
dual axis confusion



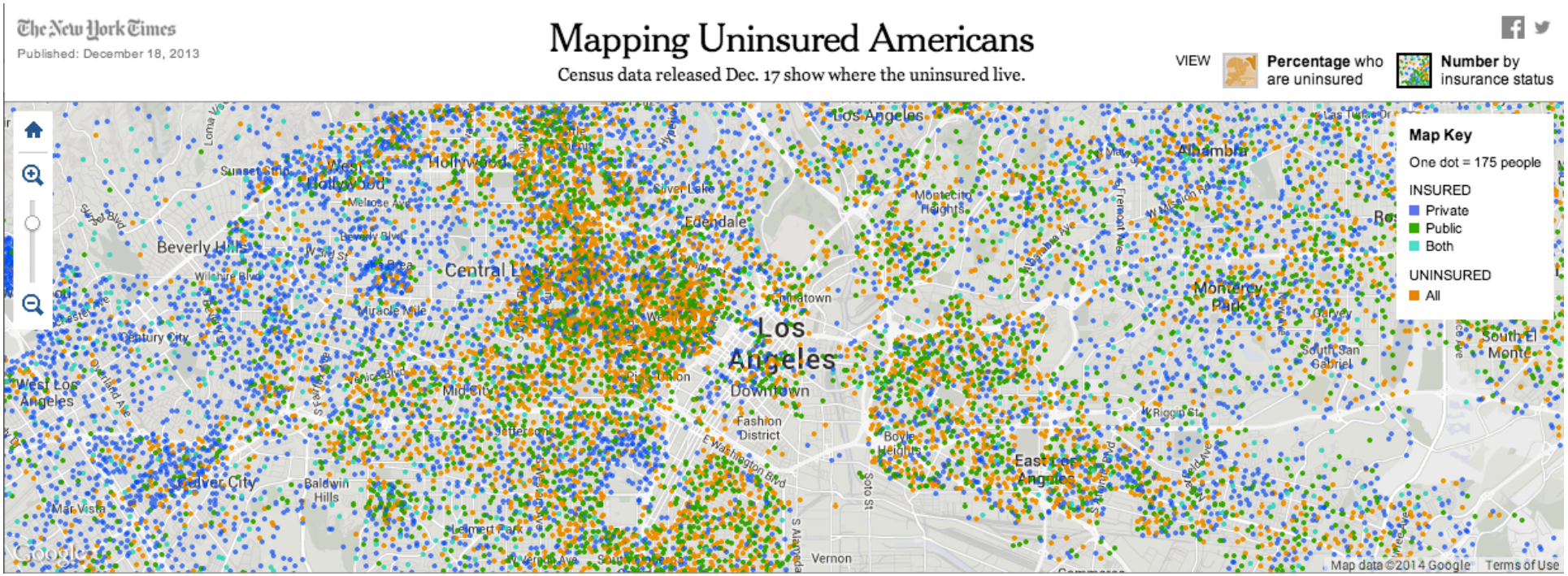
better version



difficult to read



good visualization



better visualization

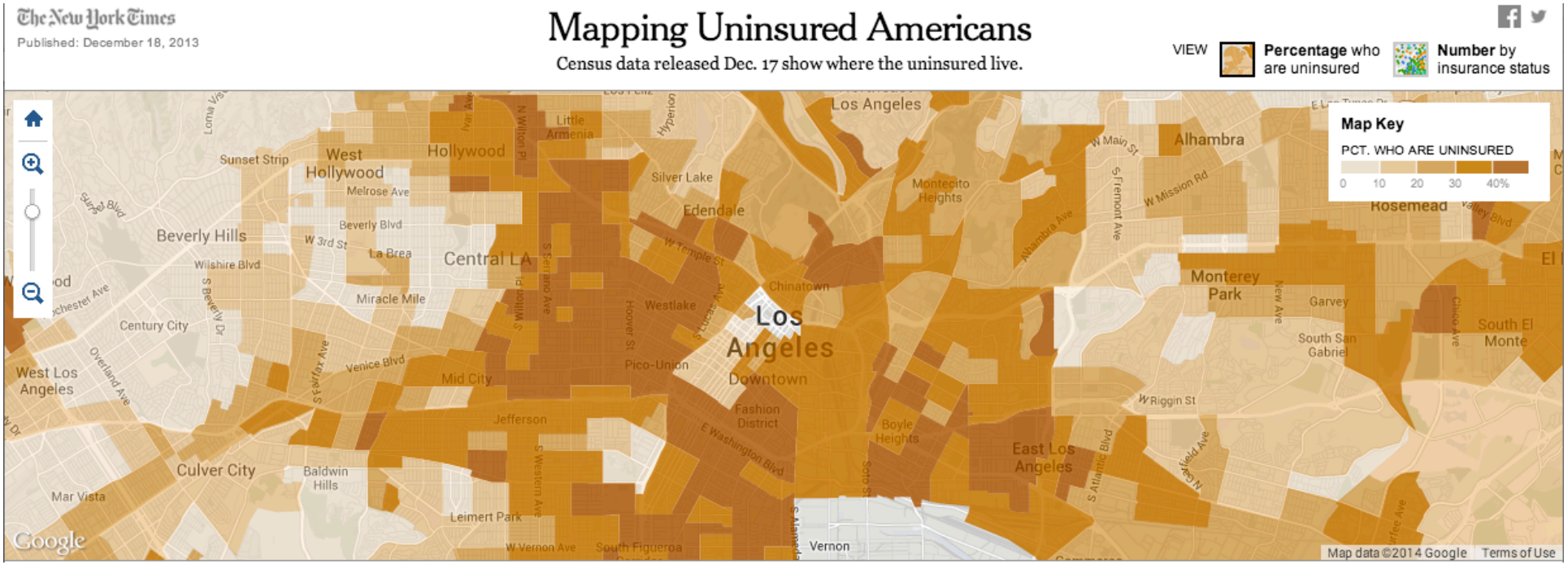
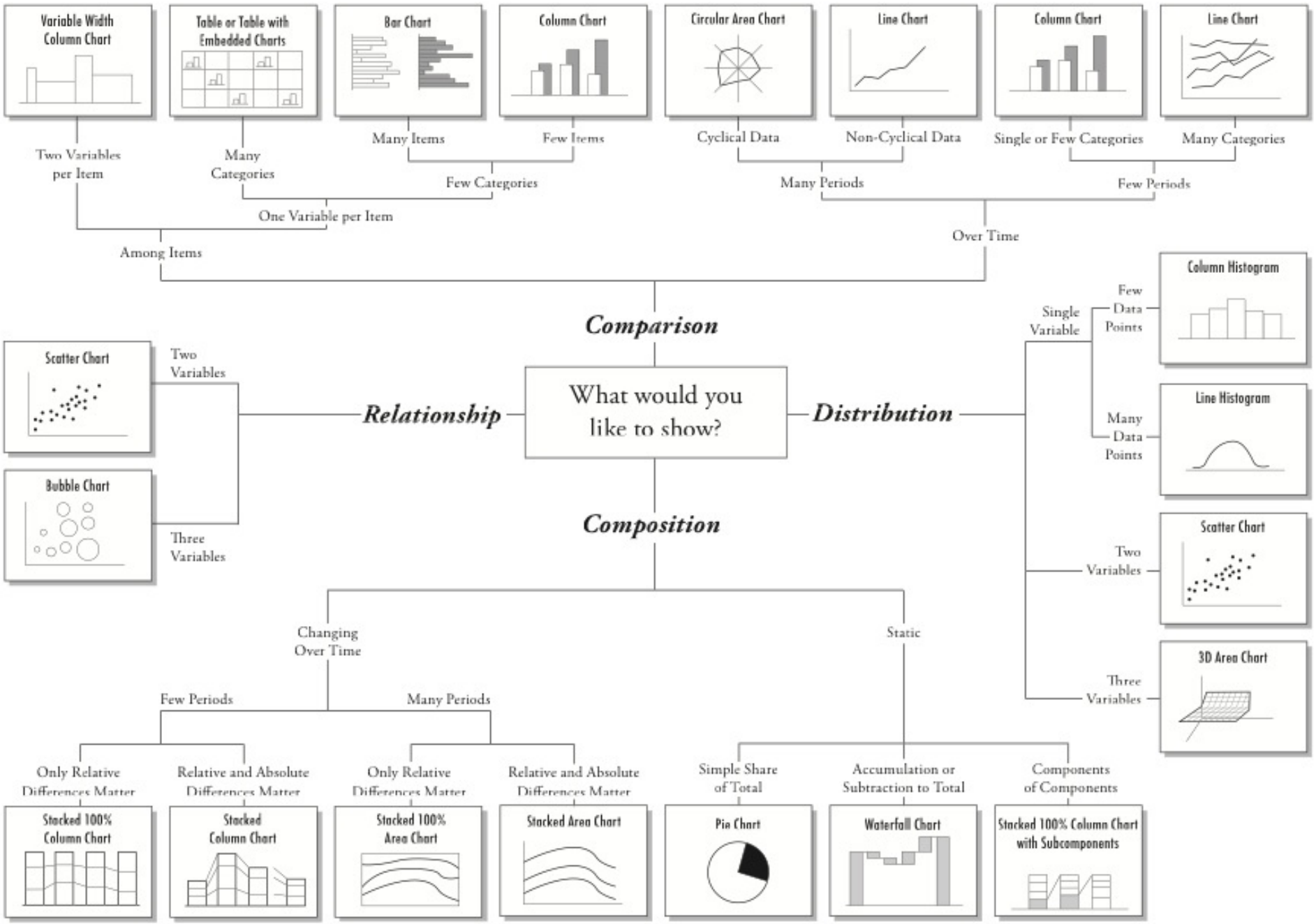


Chart Suggestions—A Thought-Starter



next steps…dynamic, interactive

references for visualization in R

R Graphics Cookbook by Winston Chang

- Same content available online
- http://www.cookbook-r.com/

docs.ggplot2.org

ggmap

- http://stat405.had.co.nz/ggmap.pdf

shiny

- http://www.rstudio.com/shiny

other visualization tools

d3.js

Tableau

network visualization

- In R: igraph, ggnet
- Other software: nodeXL, gephi, …

map visualization

- QGIS, ArcGIS, …

we have covered

ggplot's grammar of graphics

Understand your data

- scatterplot (size, shape, color, use a map as background)
- heat map
- histogram – concept of faceting

Understand your model

- add a regression line to a scatterplot
- classification output – color a map
- draw the convex hull of a set of points

Communicate

- colors
- best practices
- what not to do

More advanced

- references
- other tools
- what's next: dynamic, interactive visualizations

YOUR TURN