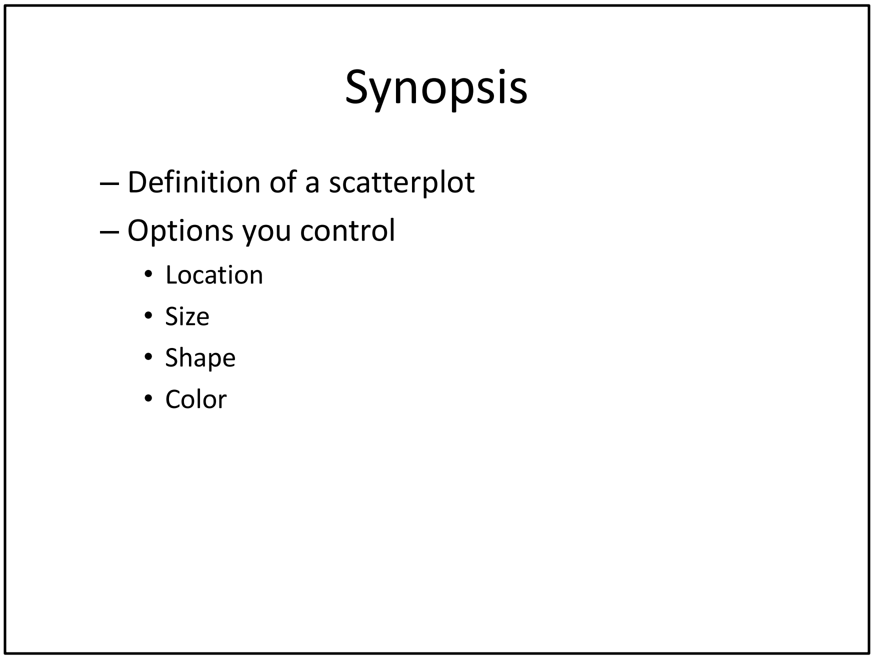
Practical suggestions for improving your scatterplots

Steve Simon



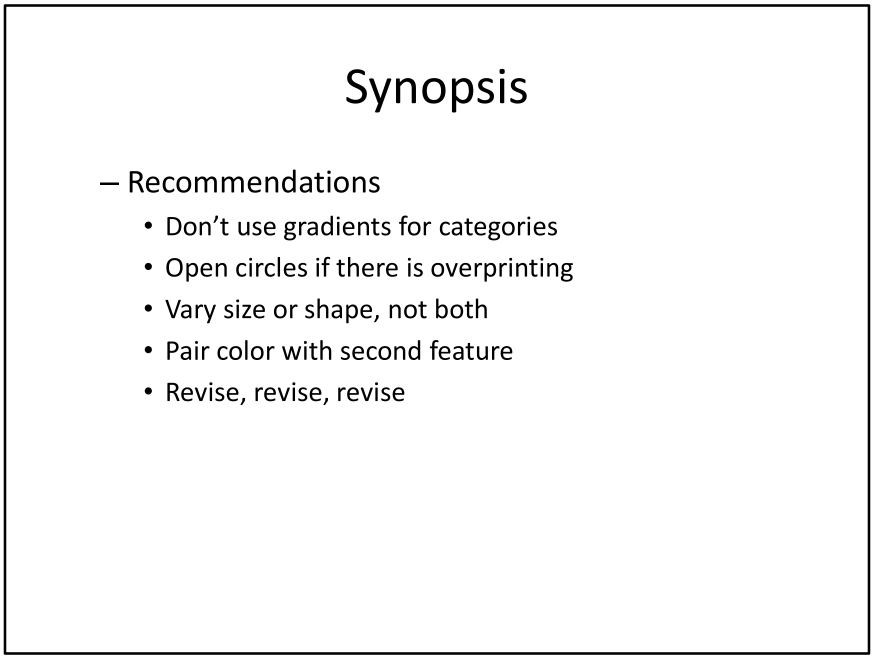
Synopsis

- Definition of a scatterplot
- Options you control
  - Location
  - Size
  - Shape
  - Color

Here is the abstract associated with this talk. I don’t want to read this word for word, but I am including it here so I can refer to it as necessary during the development of this presentation.

“Practical suggestions for improving your scatterplots”

“The scatterplot is a simple display of the relationship between two or sometime three variables. You have a wide range of options for displaying a scatterplot. In particular, you can control the location, size, shape, and color of the points in your scatterplot. Careful selection among these options will allow your audience to rapidly and accurately understand this relationship. Here are some important dos and don’ts. Don’t use a gradient to represent a nominal variable. Use open circles rather than closed circles if there is a lot of overprinting. Vary the size or the shape of your data points, but not both. Always pair color with another feature in your plots. Most importantly, never blindly accept the first graph that comes out of your software program. Revise your graphs as often as you revise your writing.”

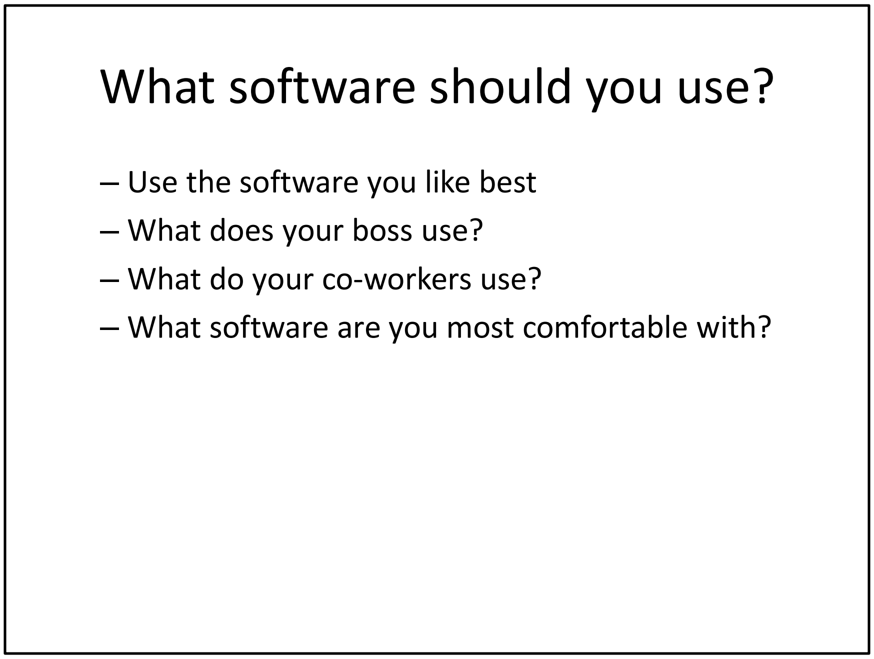


Synopsis

– Recommendations

- Don’t use gradients for categories
- Open circles if there is overprinting
- Vary size or shape, not both
- Pair color with second feature
- Revise, revise, revise

There are five general recommendations I want to make about scatterplots.



What software should you use?

– Use the software you like best
– What does your boss use?
– What do your co-workers use?
– What software are you most comfortable with?

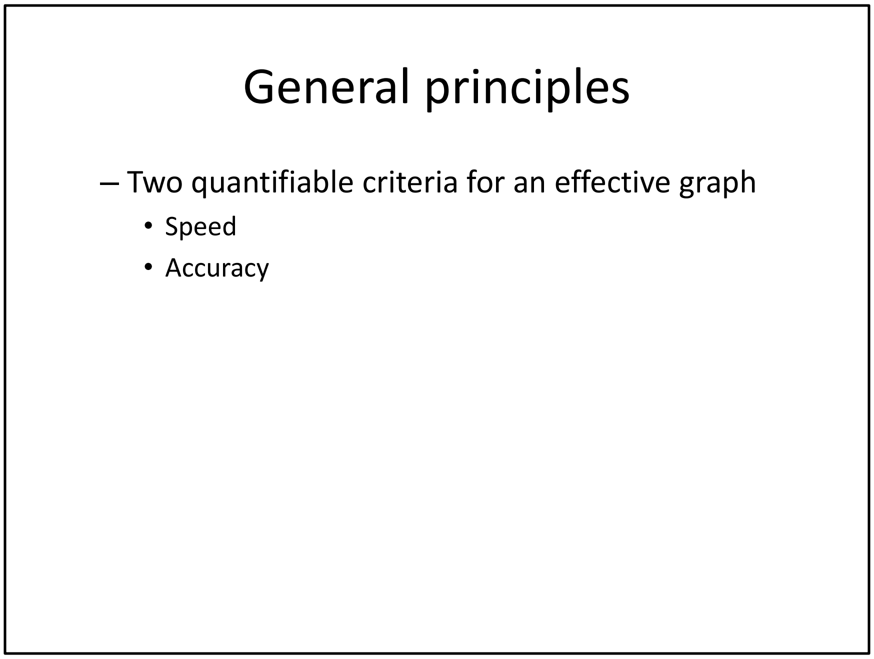
I'm a big believer in software agnosticism, and this is something that I see in the presentations by The Analysis Factor. It is a mistake to teach as if there is only one good program for data visualization.

If you are not sure what software package to use in this class, let me offer a few suggestions. First, your boss may have a strong opinion about what software that you should use. You can go to your boss and say "My teacher is a really smart guy who says that the _____ package in _____ is the best choice for data visualization." Try it and see what happens. Nothing, I suspect. One of the great tragedies in life is that the wise advice you get in this class carries very little weight in the real world.

If your boss doesn't care, see what most of your co-workers are using. They may not be as smart as I am (put on a false air of pride here) but they are a lot closer to your cubicle when this workshop ends and you have to find a quick answer.

There's also a comfort level here. Do you want a graphical user interface or a programming language. A graphical interface is great for getting work done quickly. A programming language is great for reproducibility and reusability. What fits your working style better? I don't know and it would be arrogant of me to make the presumption that I do know.

One more consideration. Some of you in this class are “ringers.” You already know visualization better than I do because you’ve been doing it for longer, with bigger and more complex data sets. You’re just here to see if I know one or two things that you don’t already know. If you’re a ringer, take the challenge of learning a new software system. It will keep you from getting too bored when I talk about all these things that you already know better than I do.



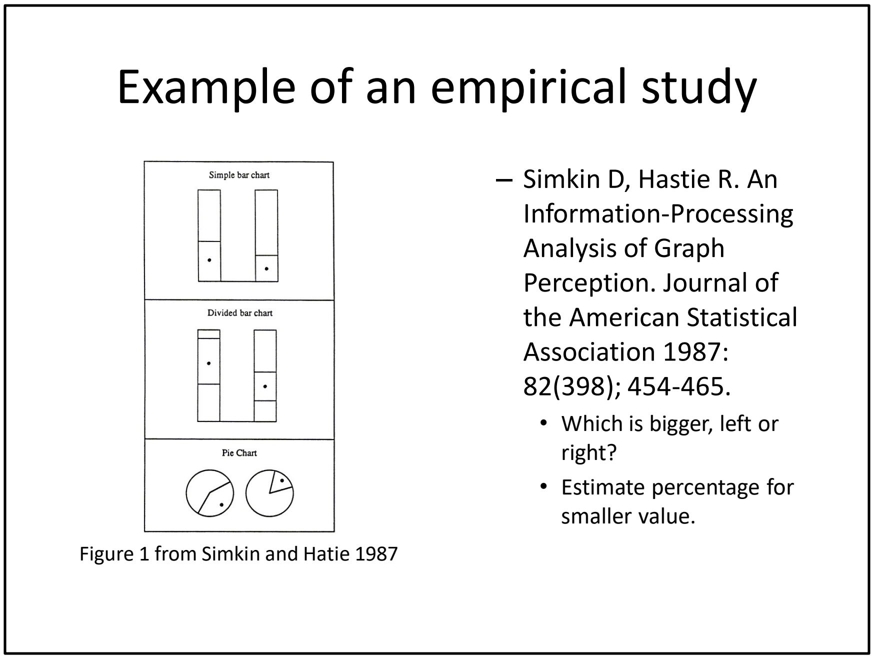
General principles

- Two quantifiable criteria for an effective graph
  - Speed
  - Accuracy

Everybody has opinions, but data trumps all. If you want to demonstrate empirically that one particular graph is more effective than another graph, you want to measure one of two things.

First, how quickly can a viewer answer a question about the graph?

Second, how accurately can a viewer answer a question about the graph?



Example of an empirical study

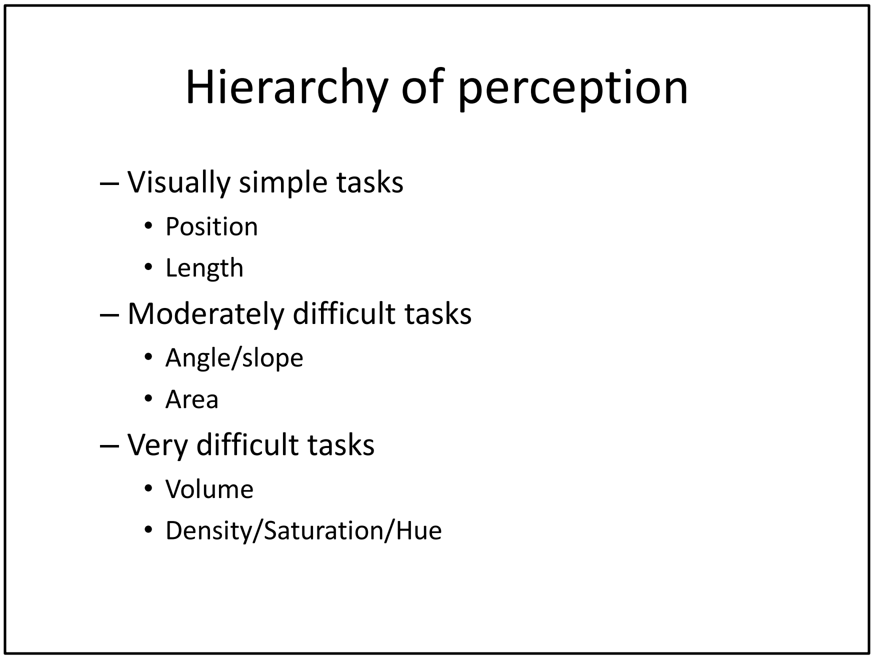
Figure 1 from Simkin and Hatie 1987

- Simkin D, Hastie R. An Information-Processing Analysis of Graph Perception. Journal of the American Statistical Association 1987: 82(398); 454-465.
  - Which is bigger, left or right?
  - Estimate percentage for smaller value.

An early example of this type of empirical study was done in 1987 by David Simkin and Reid Hastie. They showed graphs like the ones on the left, varying the size and disparity of the bars or pie wedges. They asked two questions. Looking at the the bars/wedges indicated by the dots, which is bigger the one of the left or the one on the right? What is the percentage that you would estimate for the smaller of the two?

The researchers then measured the time it took each subject to answer these questions and how accurate those answers were.

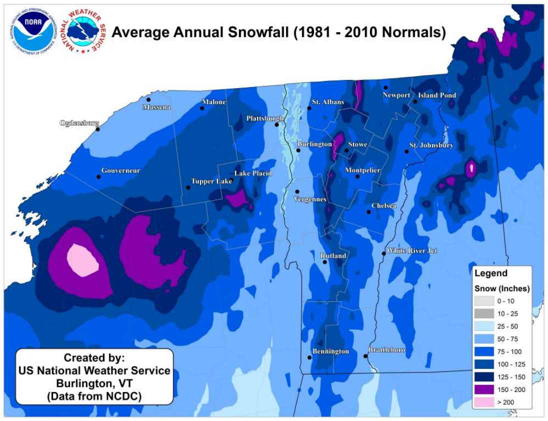
Read the paper for the full answer, but surprisingly, the pie chart turned out to be better in some settings. Better in what sense? Better in speed and accuracy.



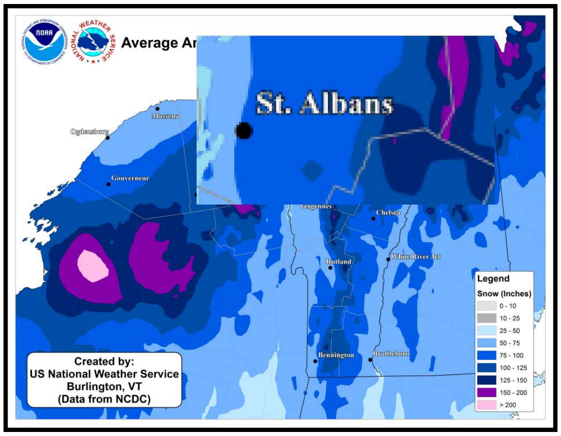
Hierarchy of perception

- Visually simple tasks
  - Position
  - Length
- Moderately difficult tasks
  - Angle/slope
  - Area
- Very difficult tasks
  - Volume
  - Density/Saturation/Hue

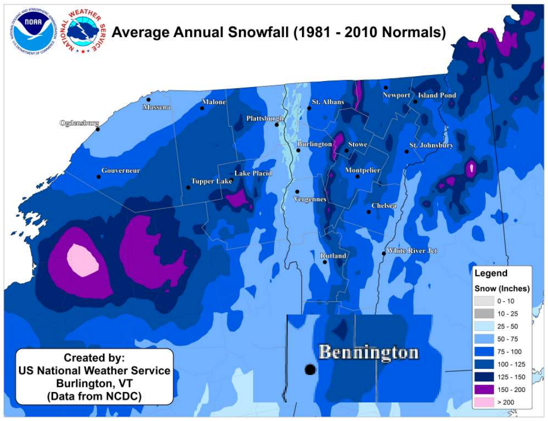
Comparison of color

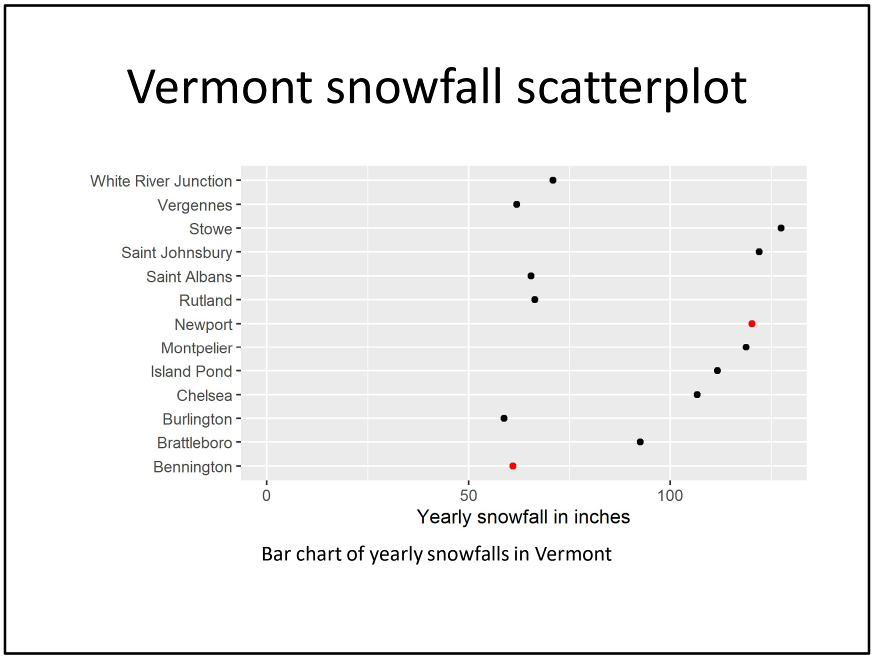


Comparison of color



Comparison of color





Vermont snowfall scatterplot

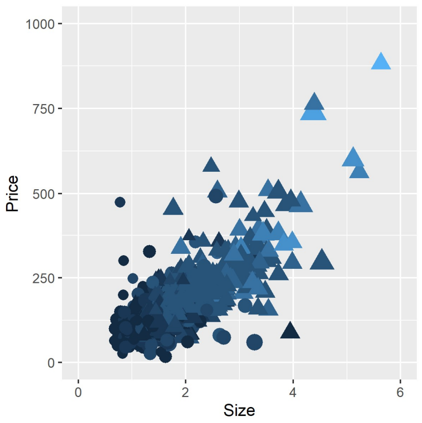
Bar chart of yearly snowfalls in Vermont

I pulled some snowfall numbers from a different website, so this data is not perfectly consistent with the maps, but it is fairly close.

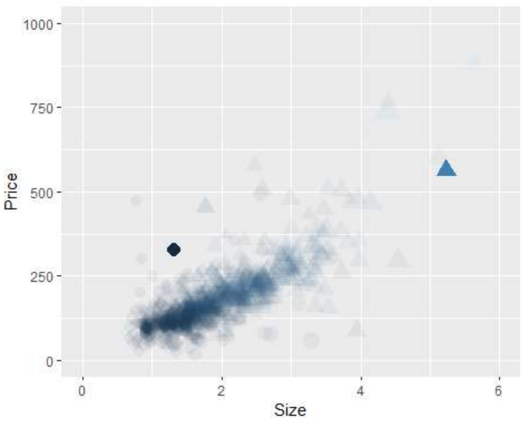
Notice how much easier it is to pick out which city has more snowfall when you display it as a scatterplot. You are judging the relative position rather than the color.

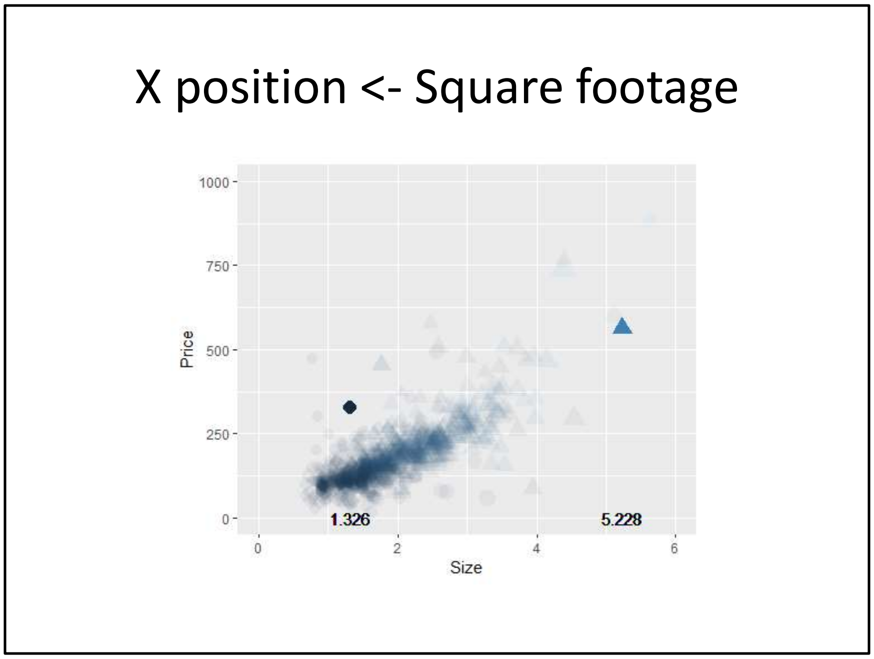
There are some questions, of course, that are still answered faster and more accurately with the map, such as whether the western edge of the state has more snowfall than the eastern edge.

A five dimensional scatterplot



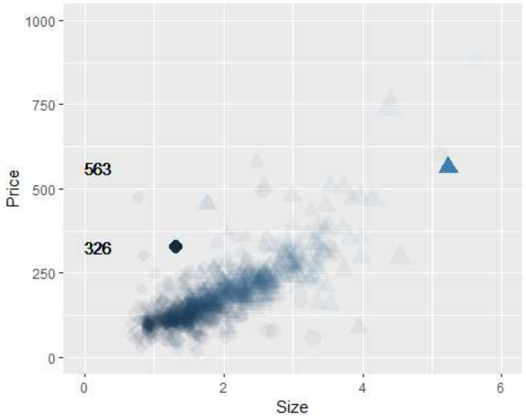
Highlight two points



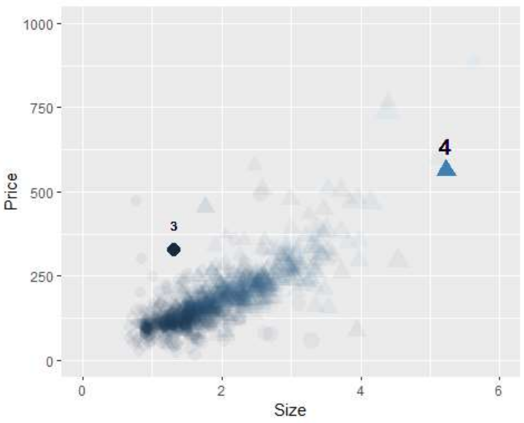


X position <- Square footage

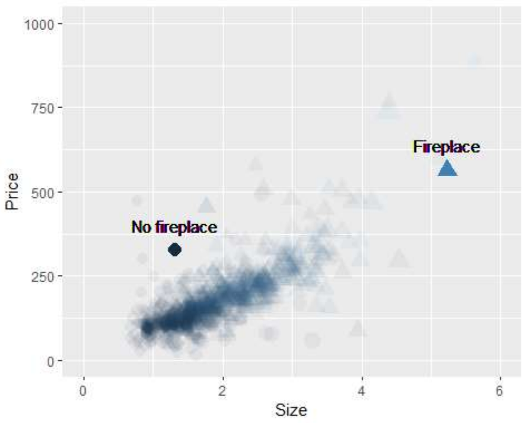
Y position <- Price



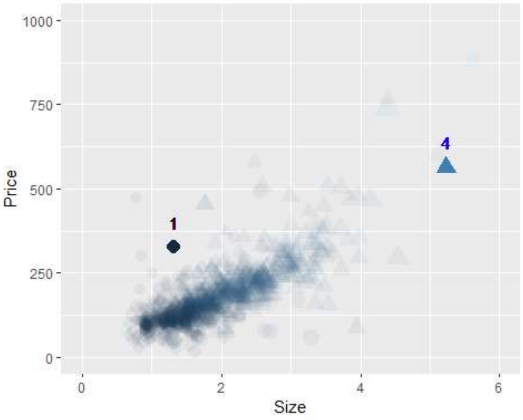
Size <- Number of bedrooms

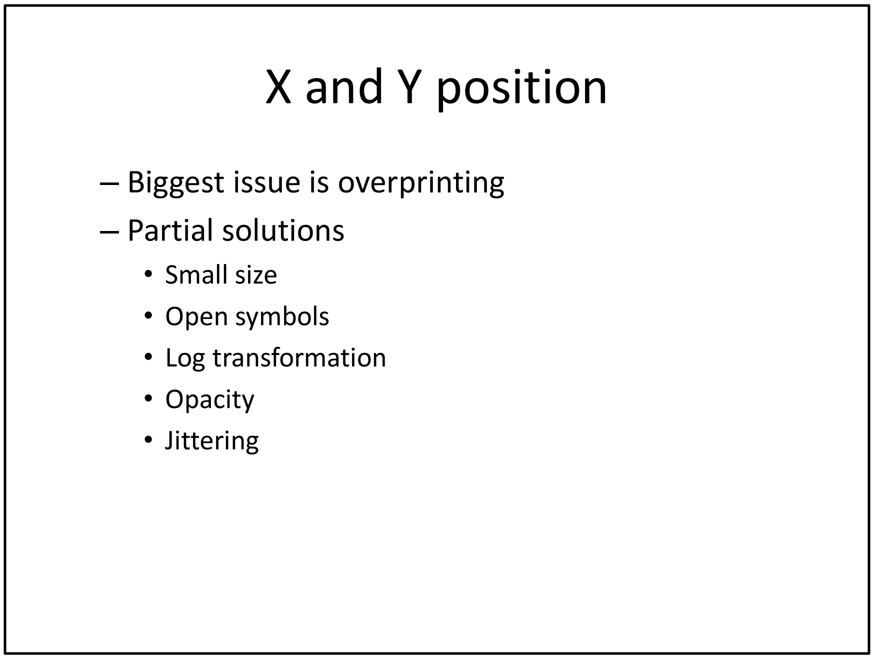


Shape <- Fireplace indicator



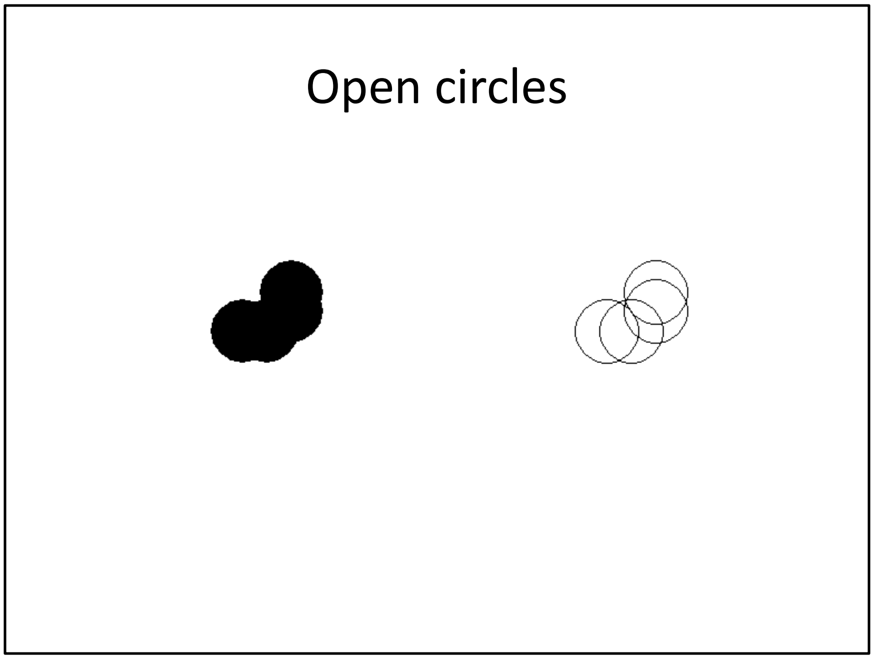
Color <- Number of bathrooms





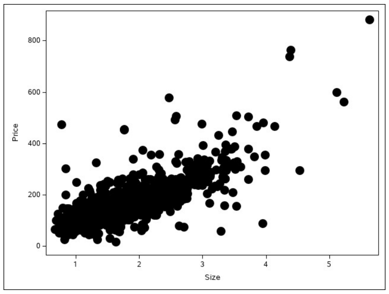
X and Y position

- Biggest issue is overprinting
- Partial solutions
  - Small size
  - Open symbols
  - Log transformation
  - Opacity
  - Jittering

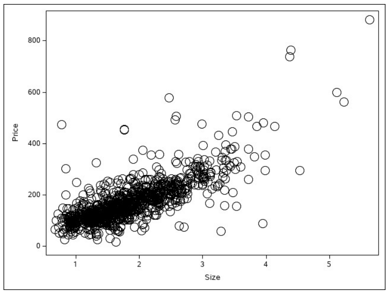


Open circles

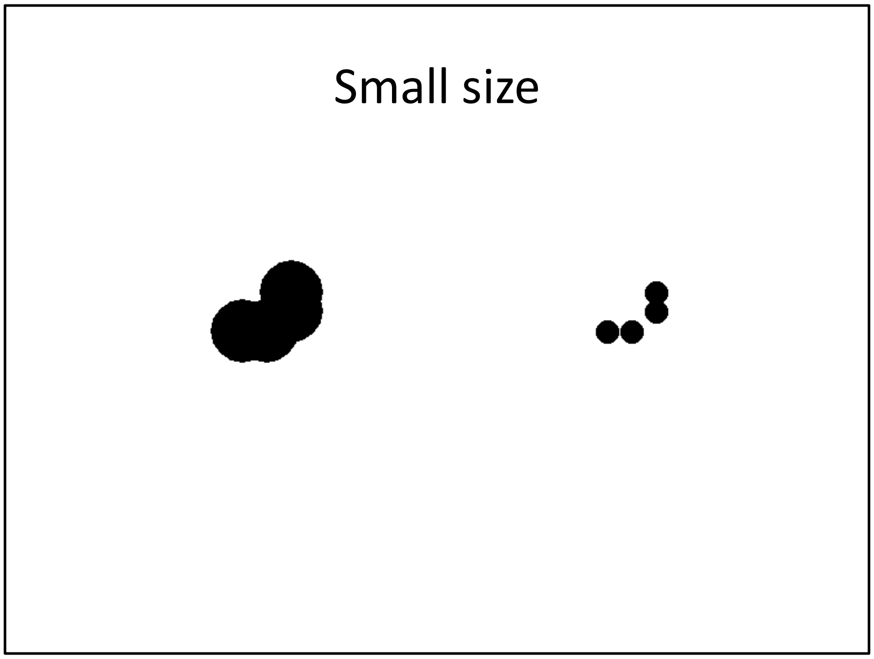
Open circles



Original scatterplot

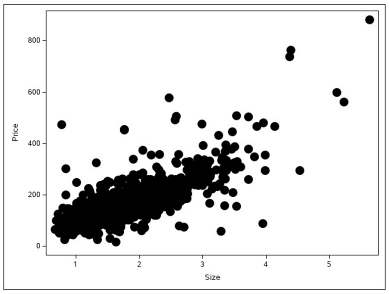


Scatterplot with open circles

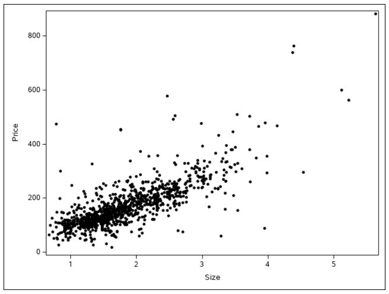


Small size

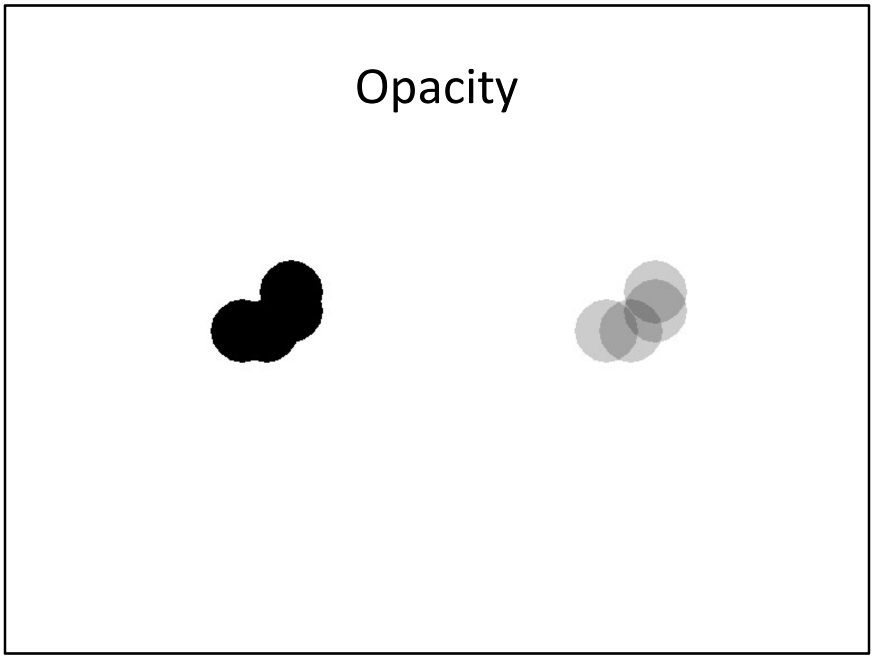
Small size



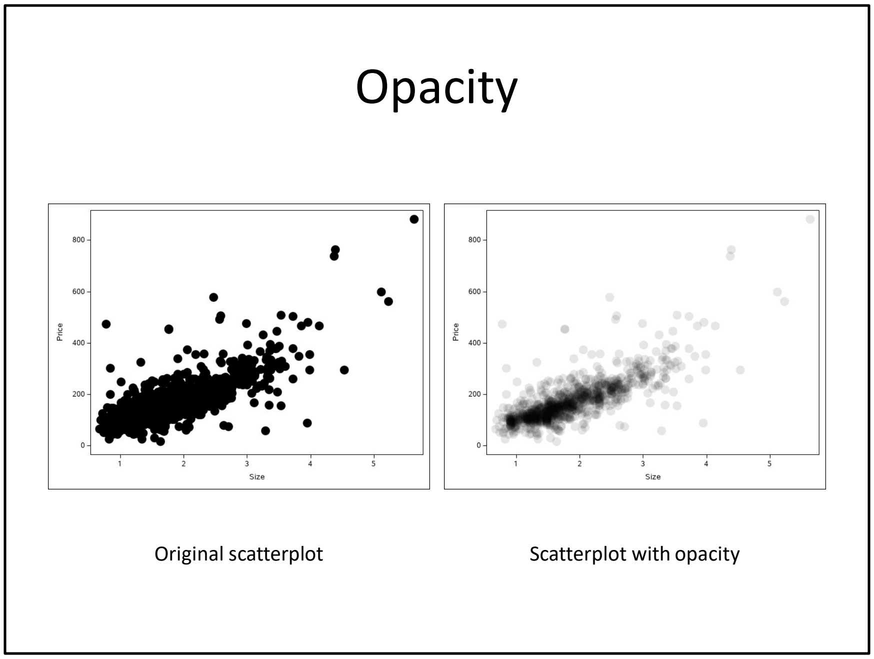
Original scatterplot



Scatterplot with smaller sized symbols



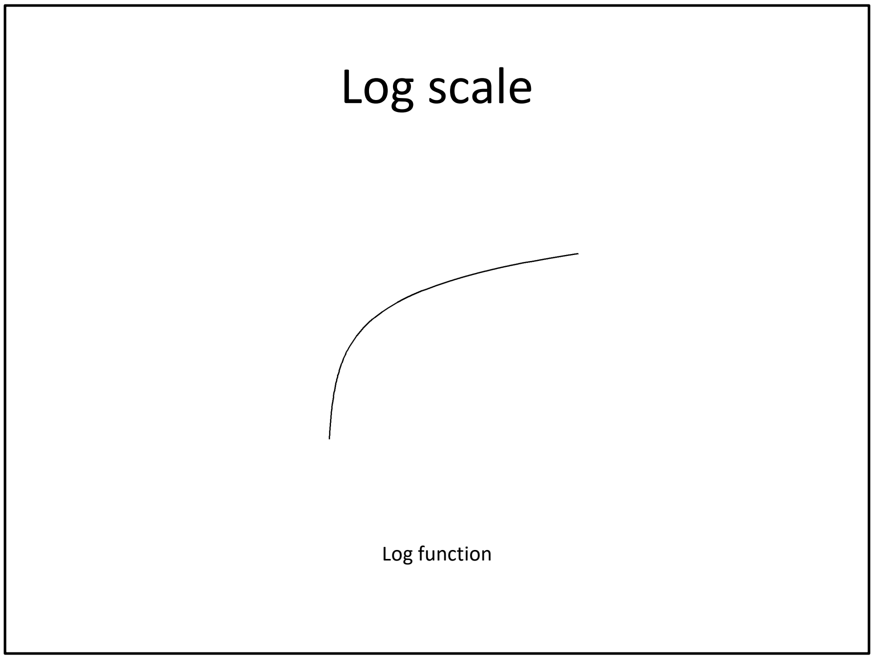
Opacity



Opacity

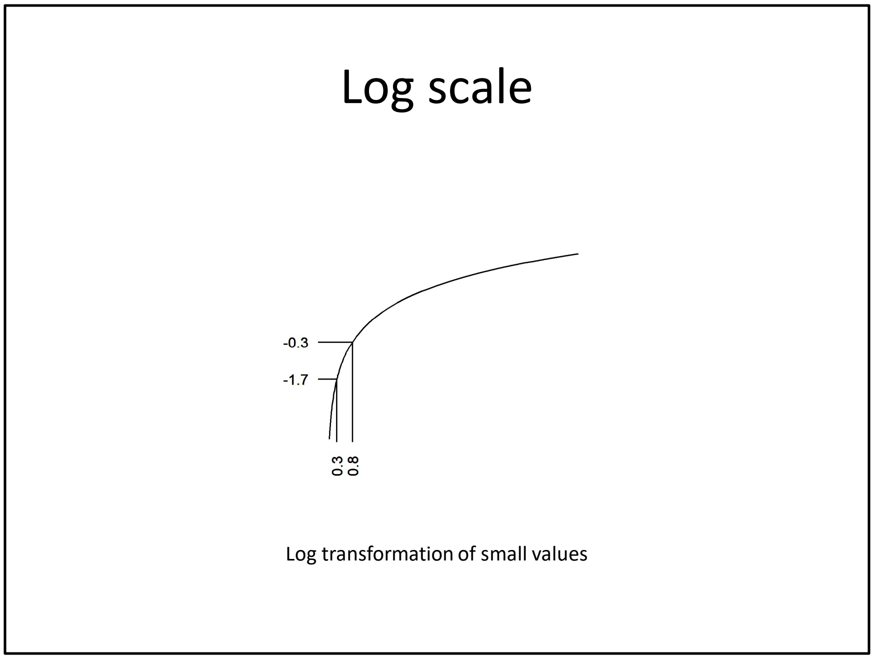
Original scatterplot

Scatterplot with opacity



Log scale

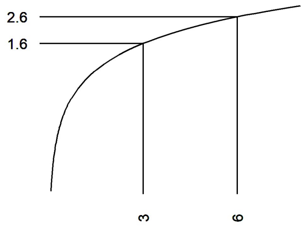
Log function



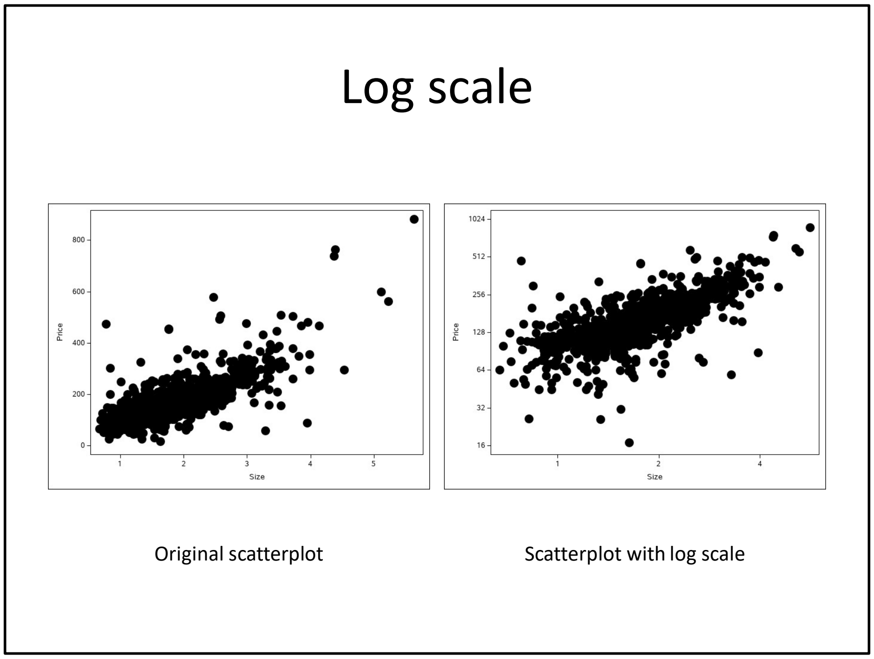
Log scale

Log transformation of small values

Log scale



Log transformation of large values

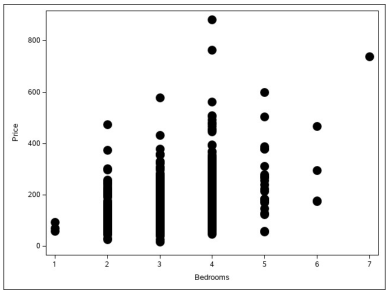


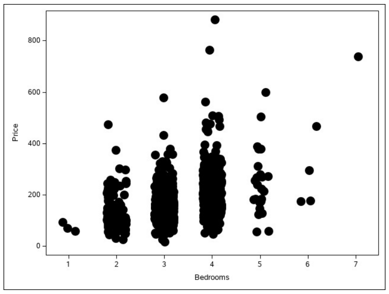
Log scale

Original scatterplot

Scatterplot with log scale

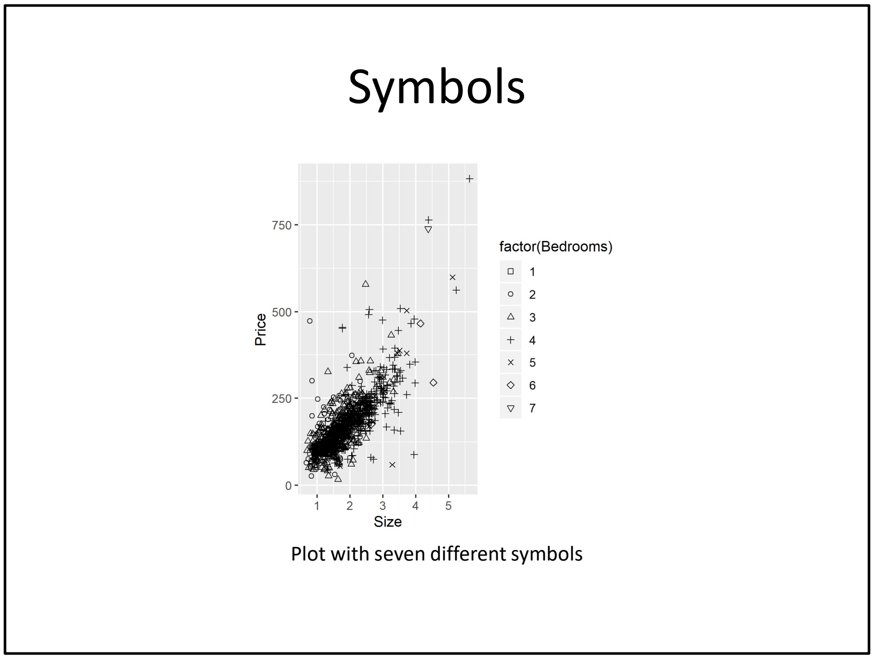
Jittering





Plot of bedrooms and price without jittering

Scatterplot with jittering



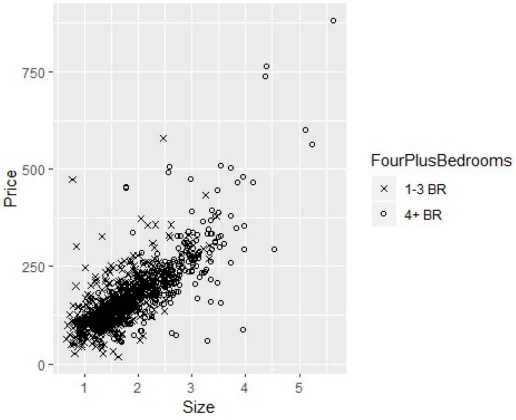
Symbols

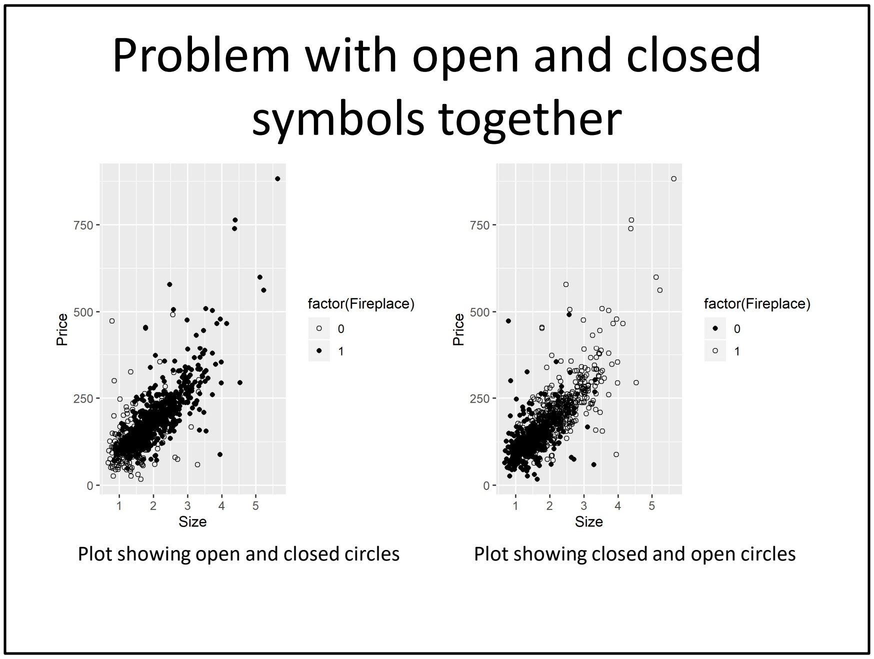
Plot with seven different symbols

This plot has two problems. First, it uses too many symbols. It turns the graph into a kind of puzzle where you are constantly going back and forth between the legend and the graph itself because no one can remember what all seven of the symbols represent.

The second problem is that number of bedrooms is not categorical. You want the greatest distinction to be between 1 bedrooms and 7 bedrooms and differences smaller than that should have proportionately less distinction.

Symbols





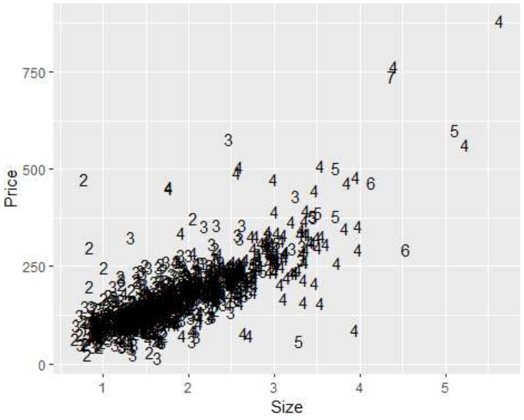
Problem with open and closed symbols together

Plot showing open and closed circles

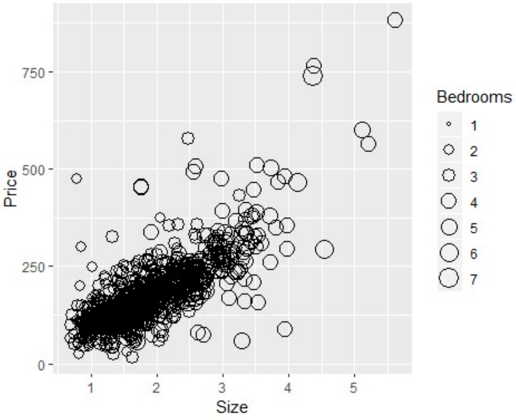
Plot showing closed and open circles

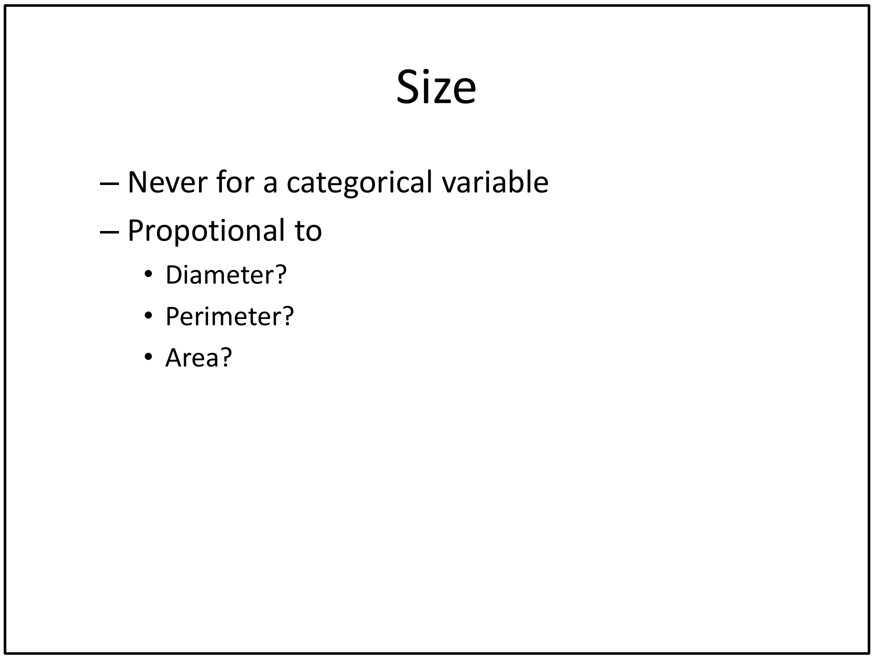
In a fight between open and closed symbols, the closed symbols always win.

Symbols



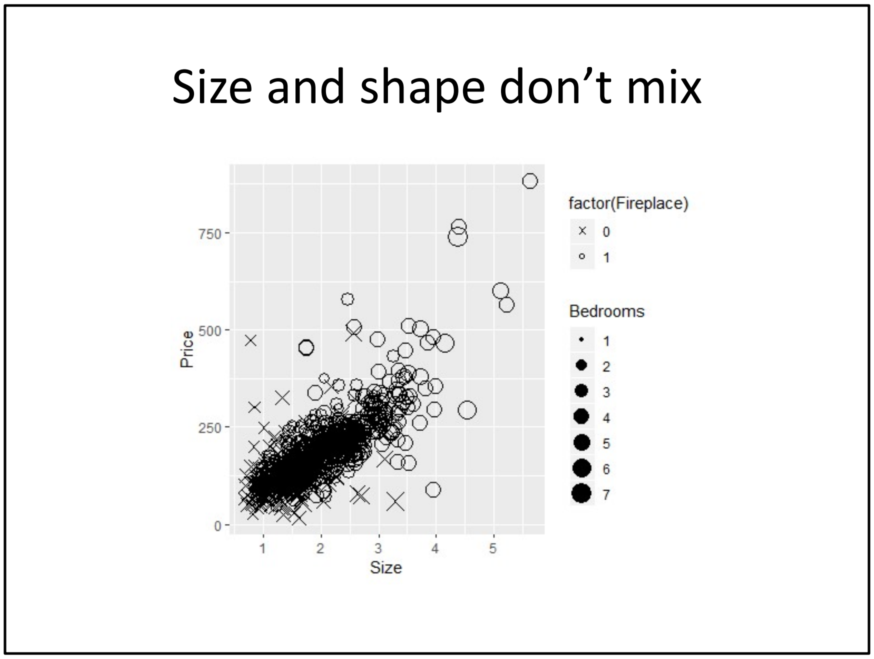
Size



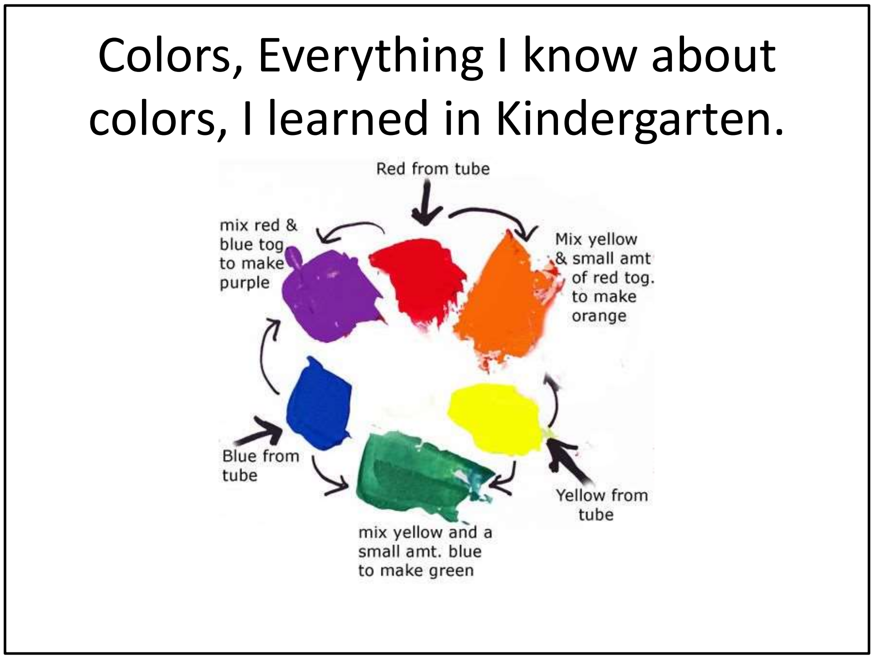


Size

- – Never for a categorical variable
- – Propotional to
  - Diameter?
  - Perimeter?
  - Area?



Size and shape don’t mix



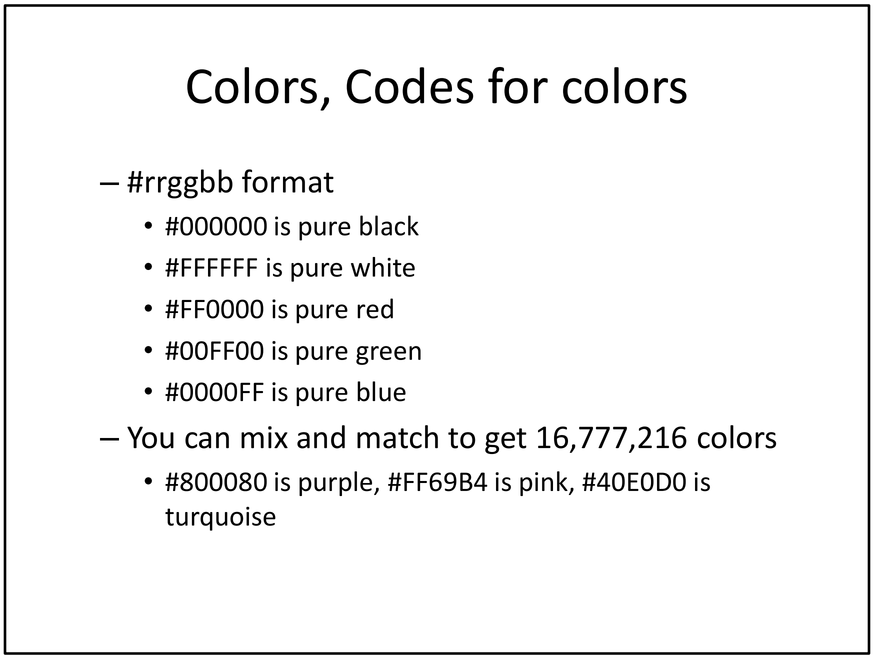
Colors, Everything I know about colors, I learned in Kindergarten.

Red from tube
mix red & blue tog to make purple
Mix yellow & small amt of red tog. to make orange
Blue from tube
Yellow from tube
mix yellow and a small amt. blue to make green

It was probably in Kindergarten where you learned the basic way to combine primary colors. Yellow plus red equals orange, Yellow plus blue equals green. Red plus blue equals purple/violet.

It doesn't work that way on a computer screen because screens use light to create colors and lights blend in different ways than paints or crayons.

Before you tackle ths computer system for colors, you need to review binary and hexadecimal number systems.



Colors, Codes for colors

- #rrggbb format
  - #000000 is pure black
  - #FFFFFF is pure white
  - #FF0000 is pure red
  - #00FF00 is pure green
  - #0000FF is pure blue
- You can mix and match to get 16,777,216 colors
  - #800080 is purple, #FF69B4 is pink, #40E0D0 is turquoise

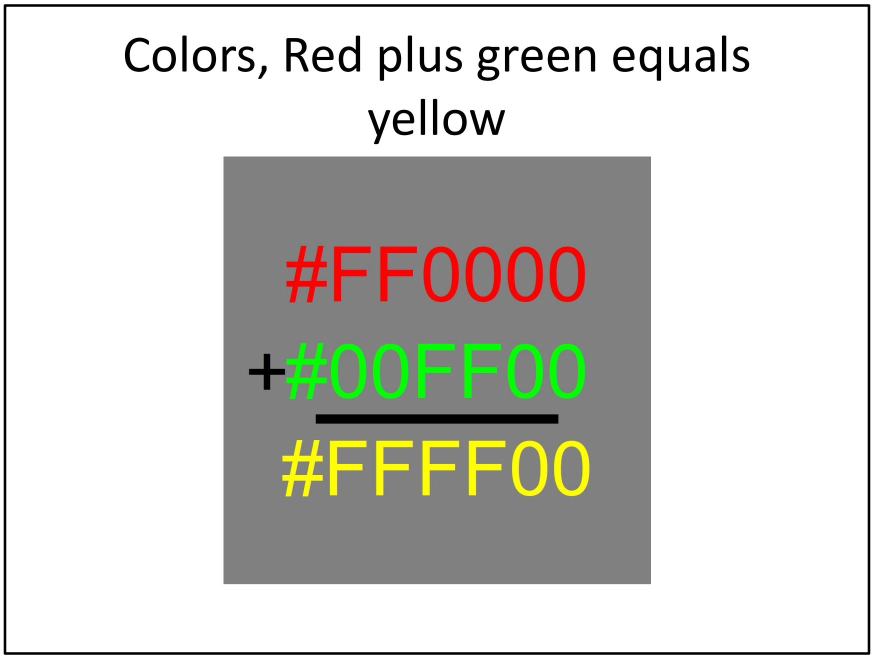
The RGB format uses six hexadecimal digits to represent colors. A hexidecimal of all zeros is pure black and at the other extreme, a hexidecimal of all F's is pure white.

The first two hexidecimal digits represent the red channel. The highest value FF for the red channel combined with zeros for the other two channels (#FF0000) equals pure red.

The next two digits represent the green channel. #00FF00, giving the maximum to the green channel and the minimum to the other two channels produces a pure green.

The last two digits represent the blue channel, and #0000FF represents pure blue.

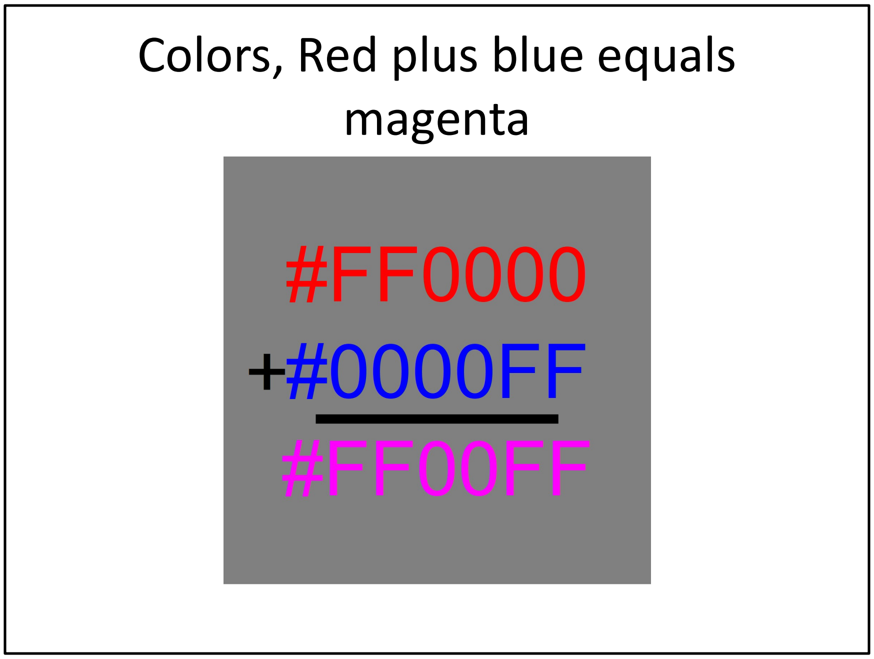
You can combine these in a variety of ways. You end up with an almost unlimited number of colors. Six hexadecimal digits allow you to produce 16^6 or 16,777,216 different colors.



Colors, Red plus green equals yellow

#FF0000
+#00FF00
#FFFF00

Whne you combine colors in the RGB system, they become lighter in color. So if you add red light (FF in the red channel) to green light (FF in the green chanel), you get yellow, which is FF in both the red and green channels.



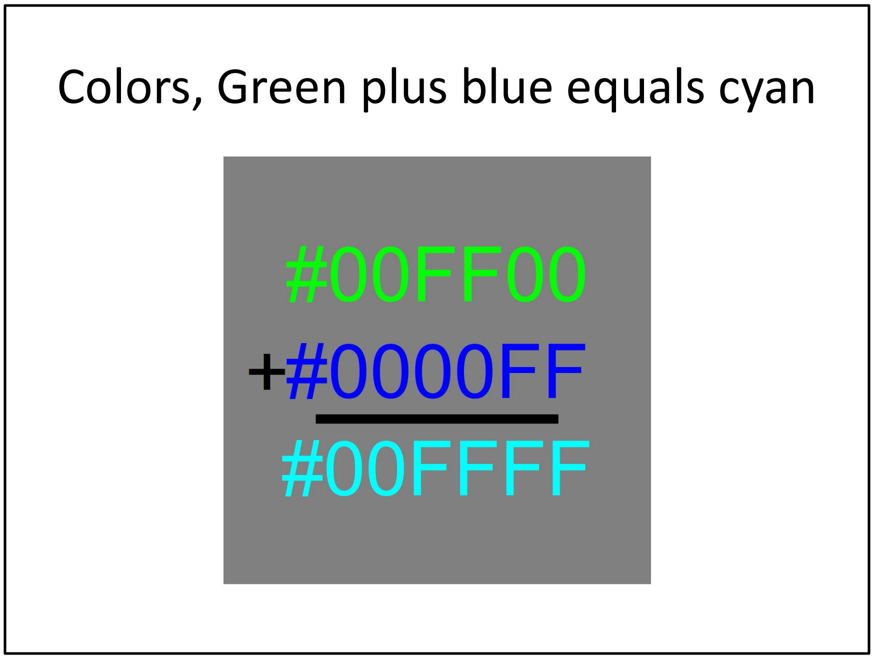
Colors, Red plus blue equals magenta

#FF0000

+#0000FF

#FF00FF

Red plus blue gives you #FF00FF, which is magenta, a light purplish red.



Colors, Green plus blue equals cyan

#00FF00

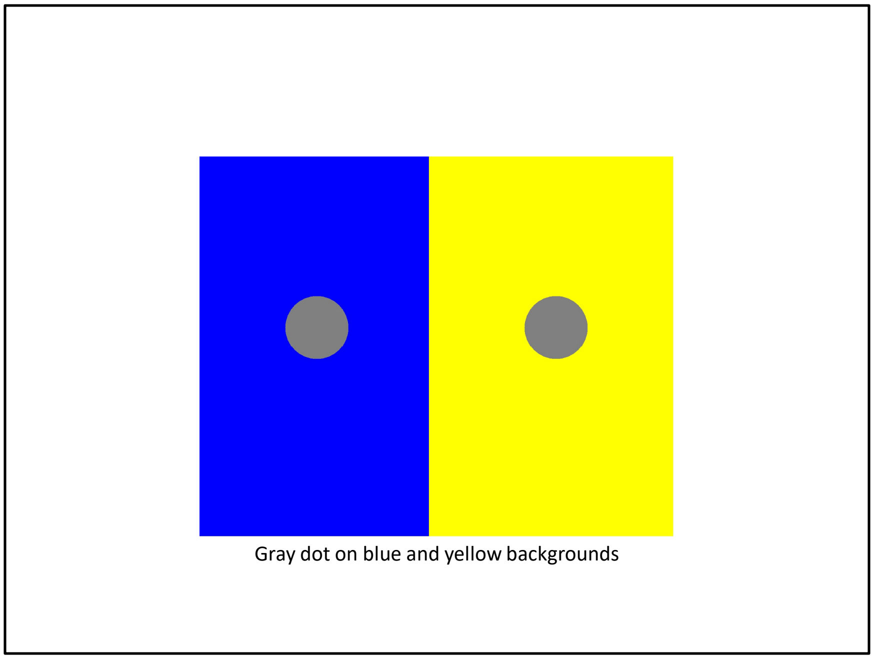
+#0000FF

#00FFFF

Green plus blue gives you #00FFFF, which is cyan, a greenish blue color.

Basic colors have different luminance

```
## Saving 5 x 4 in image
```

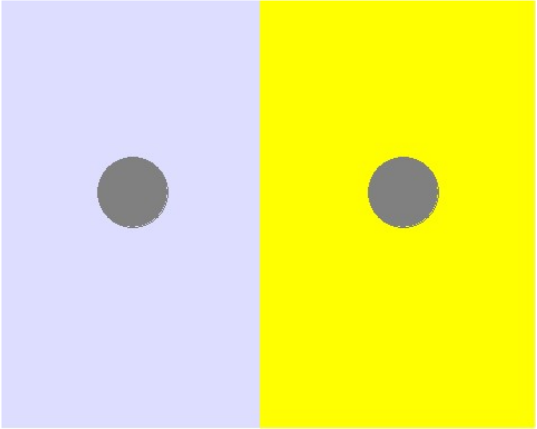


Gray dot on blue and yellow backgrounds

Among the basic colors, yellow is an outlier. It has a much higher luminance. meaning that at the same level of brightness on your computer monitor, it stimulates your optic nerves more than other colors.

This can lead to trouble. Notice the two gray dots shown on two different backgrounds. The gray in both cases is exactly halfway between black and white, but it appears darker when contrasted with yellow, because yellow has so much luminance.

Colors cause optical illusions



You can fix this by making blue lighter (closer to white).

Higher luminance colors tend to dominate a graphic image. You should try to use colors of roughly equal luminance to avoid this.

If you mix colors of different luminance, you will create artefacts that are unrelated to your data. The higher luminance colors will either tend to unfairly dominate the picture, or they will fade into the backgound and be lost.

Color

#4 TOO MANY COLORS

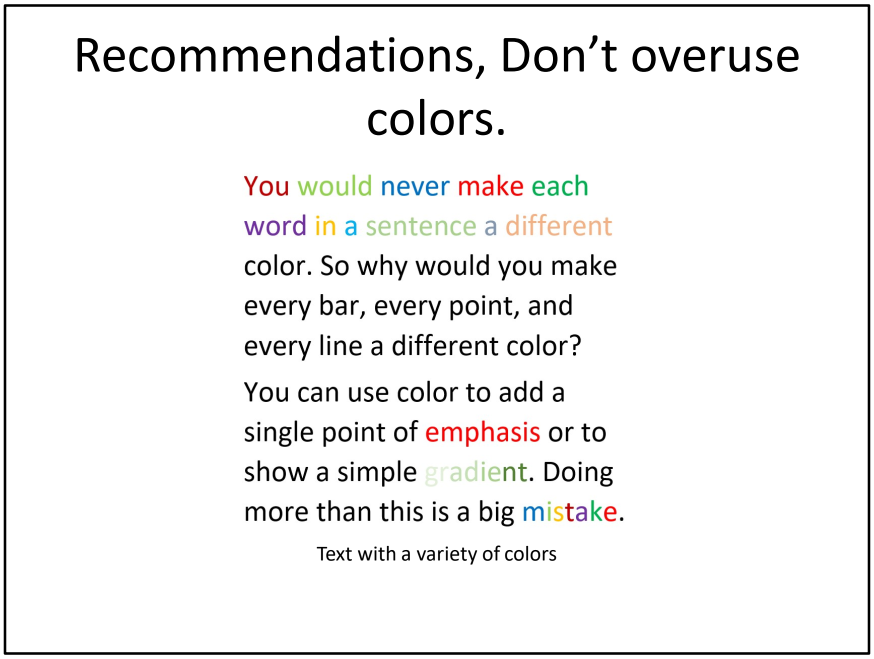
photo source: pinterest.com

Honestly, we find the best application for this quote in fashion: *"simplicity is the ultimate form of sophistication"*. Keeping it simple might be very difficult for several men, and, looking to the picture above, it really is.

Image of man wearing three bold colors

It's a well known fasion mistake to wear too many colors at the same time. Maybe this guy could get away with it, but most of us would look like idiots if we tried to dress that way.

There's a similar lesson for data visualization.



Recommendations, Don’t overuse colors.

You would never make each word in a sentence a different color. So why would you make every bar, every point, and every line a different color?

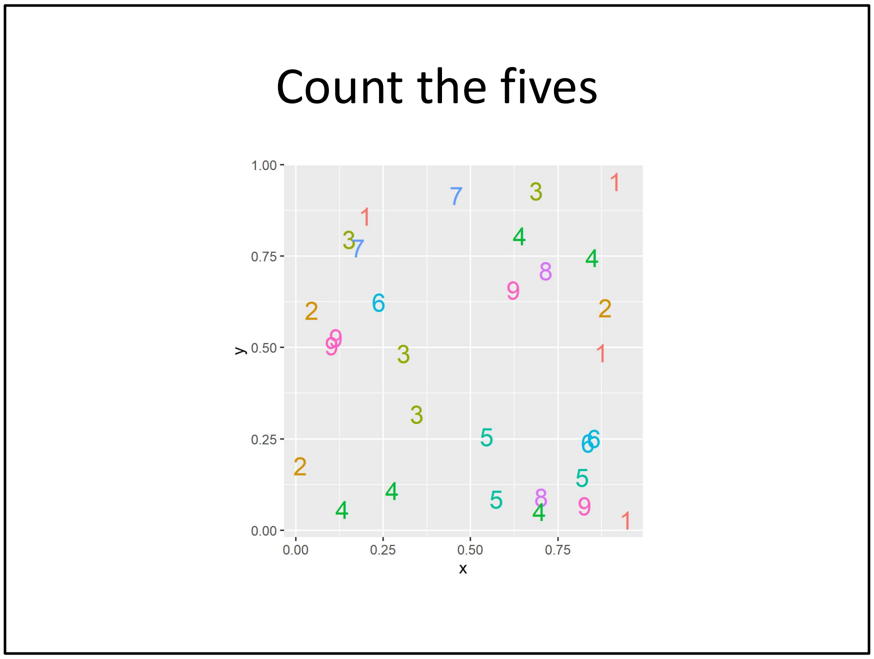
You can use color to add a single point of emphasis or to show a simple gradient. Doing more than this is a big mistake.

Text with a variety of colors

Naomi Robbins, an expert on data visualization, made an interesting observation. You would never make each word in a sentence a different color. So why would you make every bar, every point, and every line a different color?

Too many colors dilutes the impact that color can have.

You can use a second color to add emphasis. Or maybe a gradient between two different colors could work. Doing more than this is usually a big mistake.

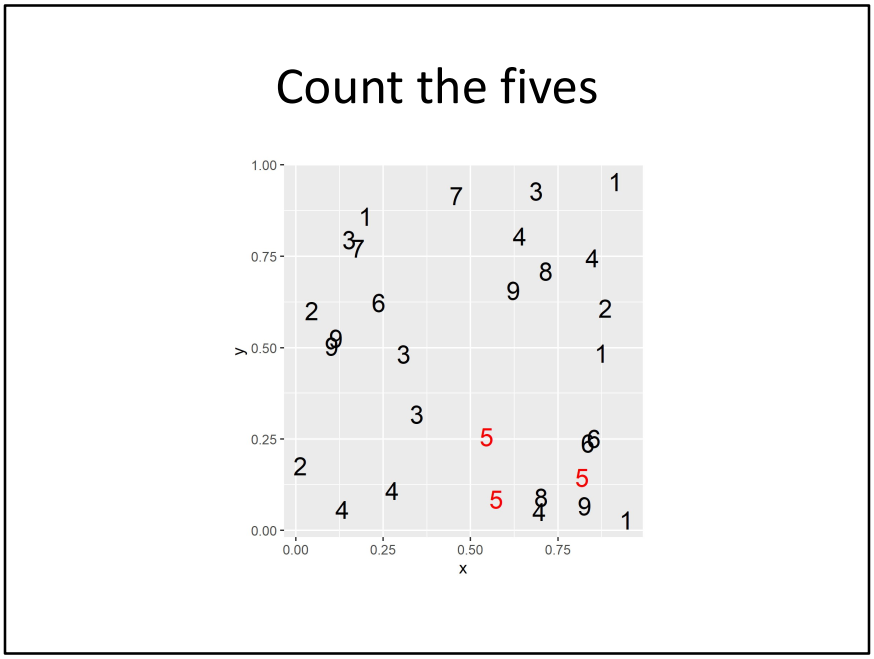


Count the fives

Here’s an exercise that adapted from Olson and Bergen.

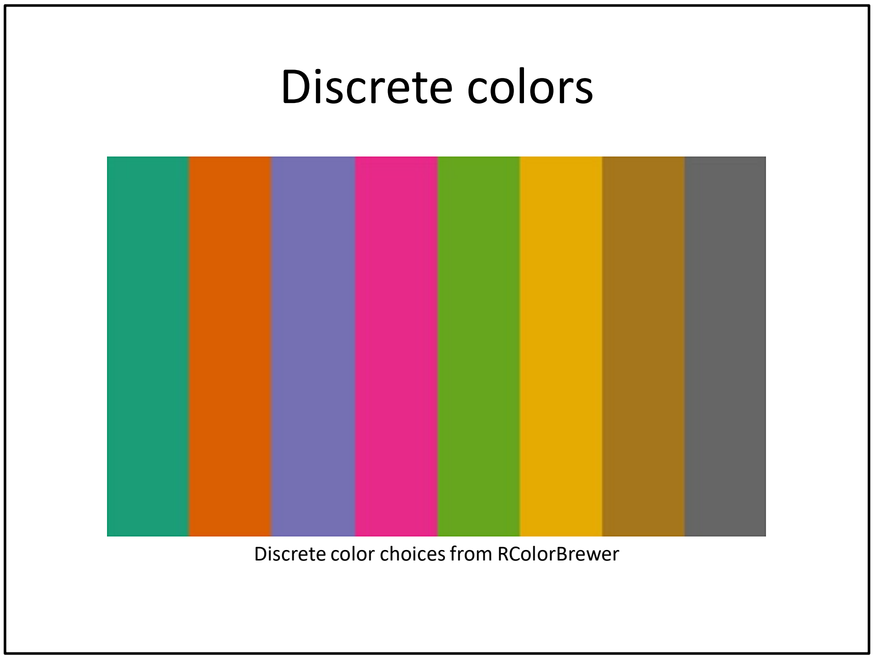
How many fives are there in this picture. I’ve used a different color for each number to make it easier for you to pick out any particular number. It takes a while, but you can see that there are three 5’s, clustered in the lower right corner of the graph.

Did the colors help? Well, not all that much. It is hard to pick out nine colors and not have a few of them look very similar. In particular, the 5’s and the 6’s are pretty close, as are the 8’s and the 9’s.



Count the fives

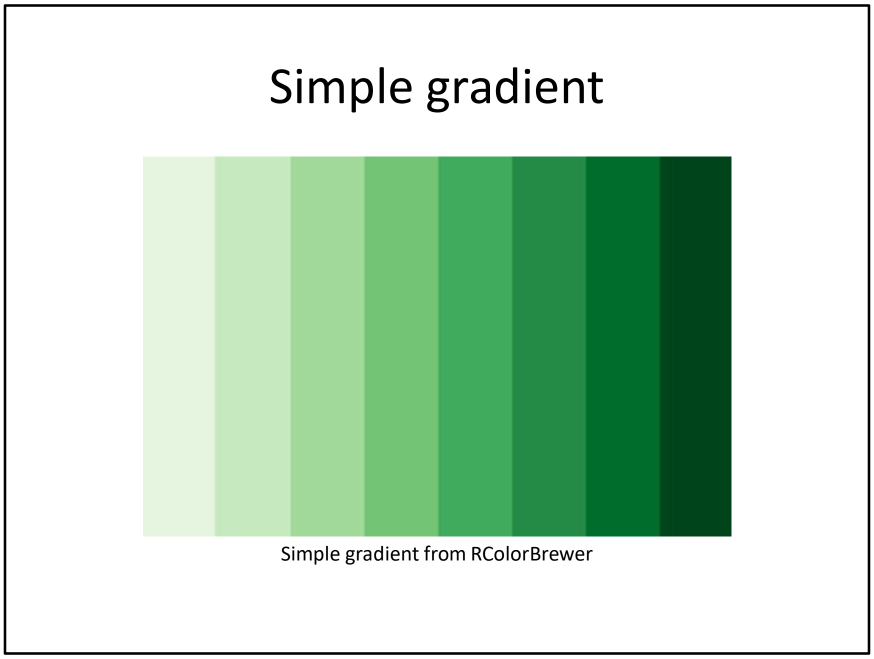
When you use a bit of restraint and only show two colors, you make the process of identifying all the fives much easier.



Discrete colors

Discrete color choices from RColorBrewer

This is a nice set of colors. Each color is distinct from each other color and they are all roughly the same level of luminance. This makes the most sense for categorical data.

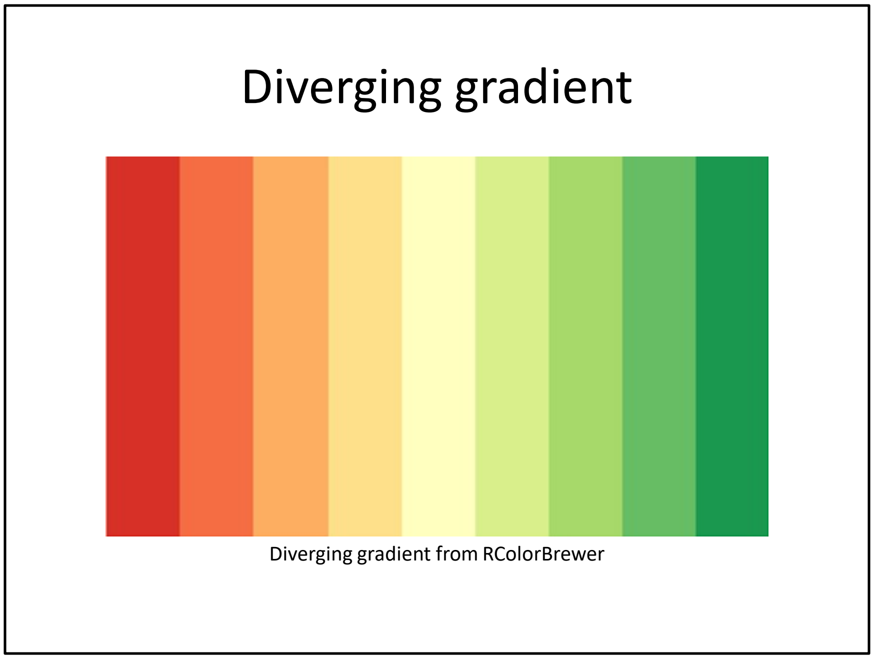


Simple gradient

Simple gradient from RColorBrewer

A simple gradient transitions within a single color, usually from lighter shades of that color to darker shades.

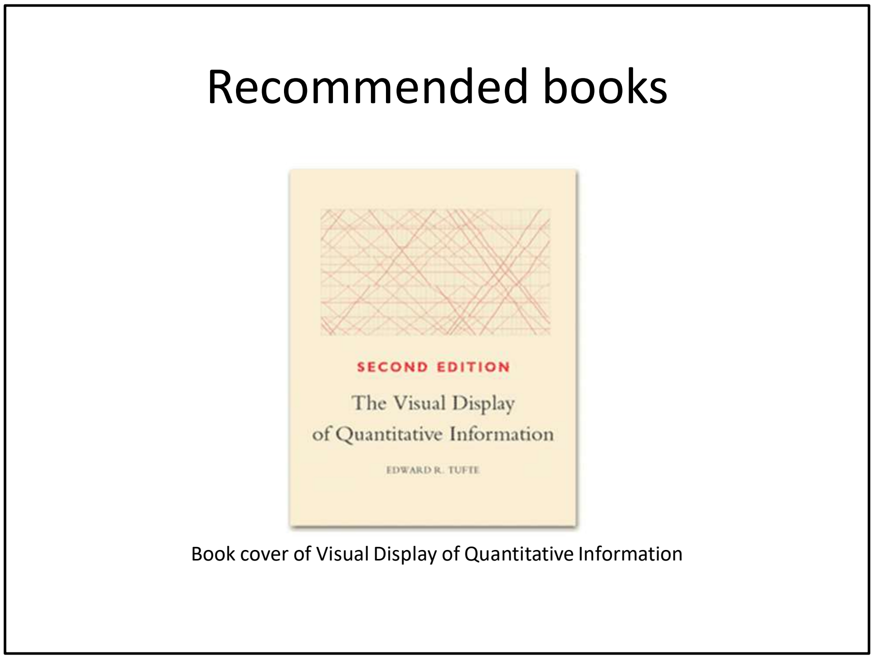
Depending on the nature of your plot, one end of the gradient will be emphasized and one end will be de-emphasized.



Diverging gradient

Diverging gradient from RColorBrewer

A diverging gradient moves between two distinct colors and takes a side trip in the middle to a third color. Typically, the middle color in a diverging gradient has a much higher luminence or a much lower luminence than the two extremes and is intended to fade into the background. The diverging gradient tends to emphasize the extremes and de-emphasize the middle.

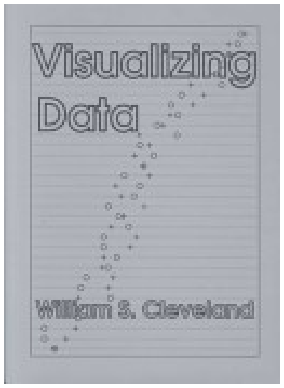


Recommended books

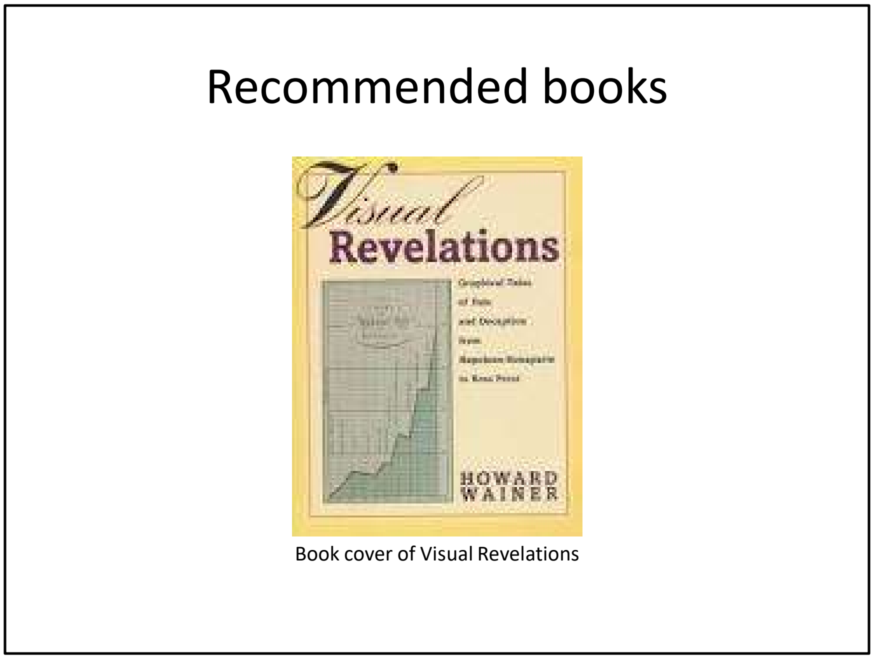
Book cover of Visual Display of Quantitative Information

Edward Tufte is a very interesting person. Very opinionated, and right most of the time. When his book suffers, it is from an attention to a guiding principle that is so rigid that it misses out on the times when there are exceptions to every principle. This is not the book to start with, but one that you should read after you’ve been doing visualization for a while. It will help you develop an eye for what works and what doesn’t work.

Recommended books



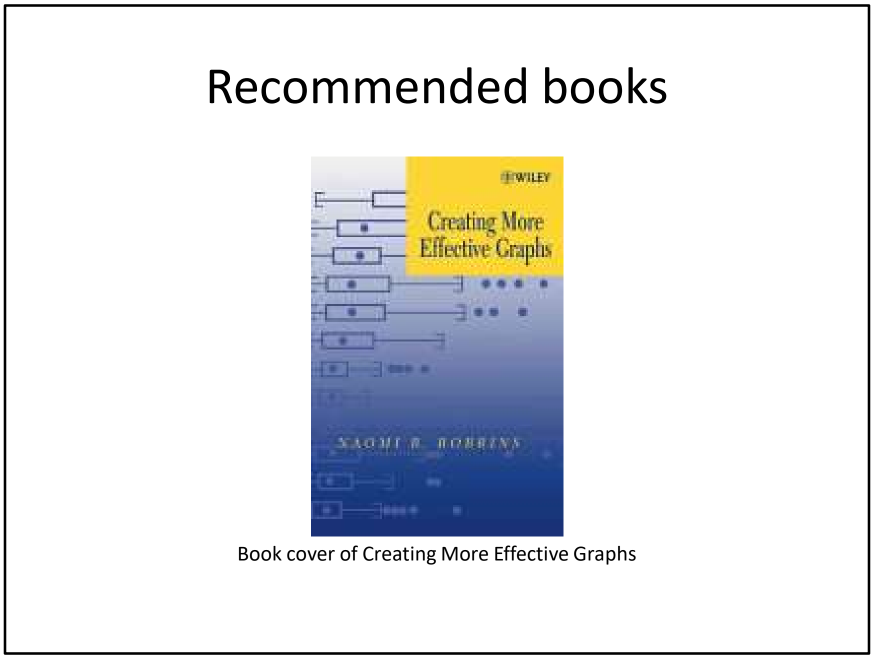
Book cover of Visualizing Data



Recommended books

Book cover of Visual Revelations

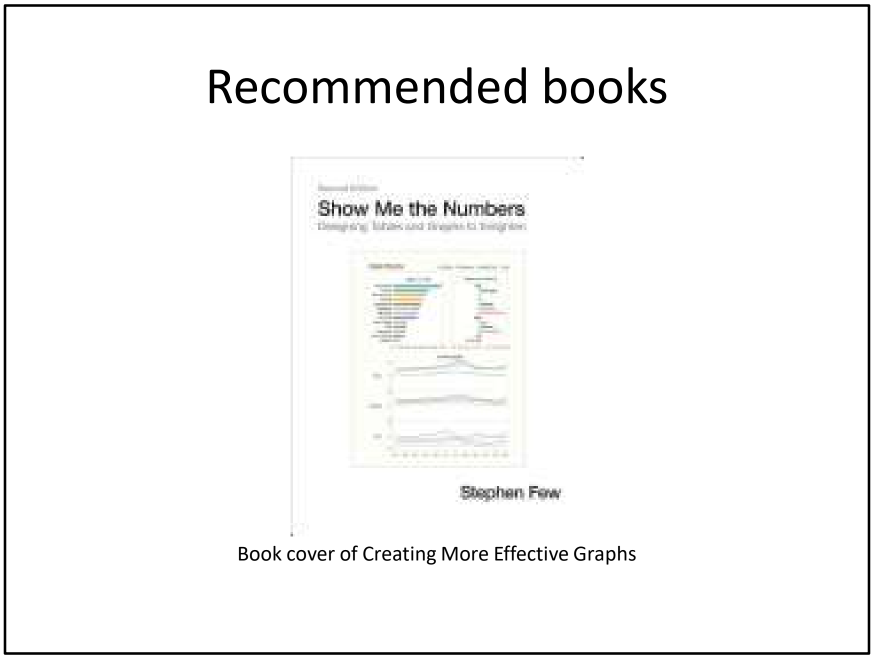
I learned more from this book than any other. It might be a bit dated today, but it helps you understand why a graph works from the underlying issues of the psychology of perception. This is also a good book to read if you want to see some of the pioneering changes that were made in data visualization just before the turn of the century. If you are just starting out in data visualization, this is not going to be your first book.



Recommended books

Book cover of Creating More Effective Graphs

While the previous two books were more theoretical, this book is very applied and is an excellent starter book for anyone wanting to learn more about data visualization.



Recommended books

Book cover of Creating More Effective Graphs

This is another very applied book. Both are equally good, but Robbins book seems a bit friendlier to me.