

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 marks 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 marks, 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. If you make a choice that makes me happy and your boss mad, I won't be able to get you that promotion you've been hoping for.

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.

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

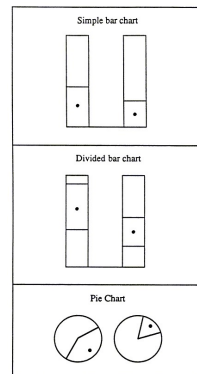


Figure 1 from Simkin and Hastie 1987

— 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?

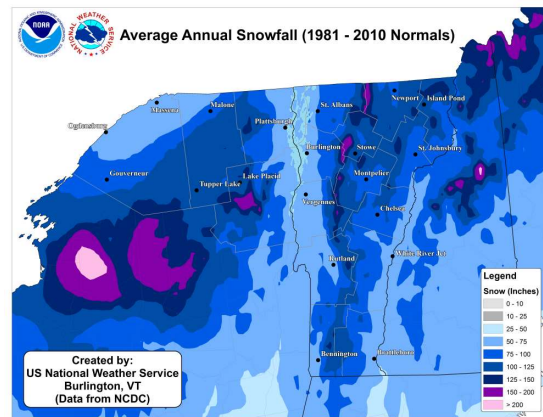
The researchers then measured the time it took each subject to answer these questions and how accurate those answers were.

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

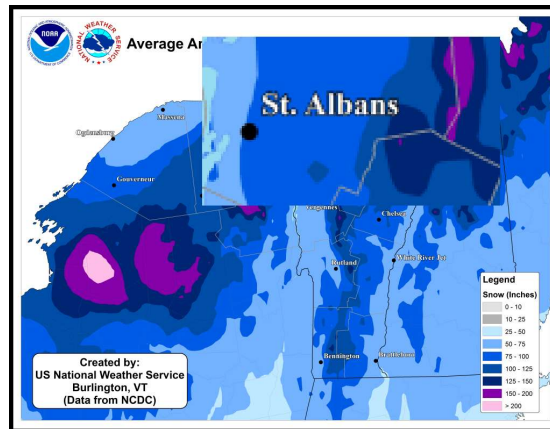
- Visually simple tasks
 - Position
 - Length
- Moderately difficult tasks
 - Angle/slope
 - Area
- Very difficult tasks
 - Volume
 - Density/Saturation/Hue

Comparison of color



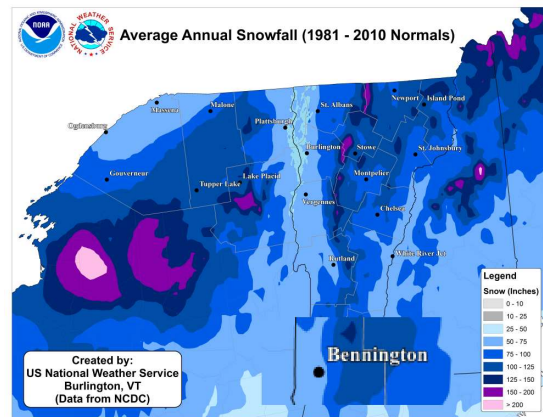
This map uses color to show average yearly snowfall in Vermont and parts of New York. One question you might ask is how much does it snow in one Vermont city versus another.

Comparison of color



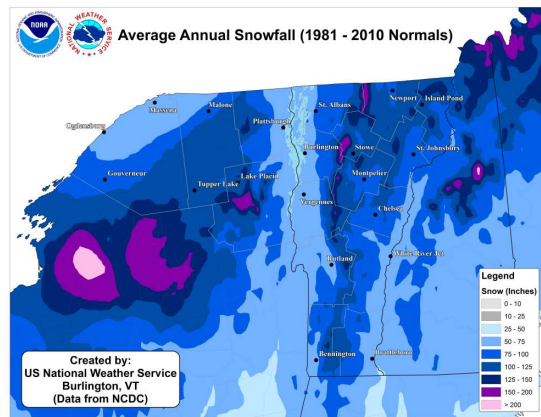
I've magnified a northern part of the state to highlight St. Albans.

Comparison of color



Here's a city in the southern part of the state, Bennington. Okay, keep those two states in mind, now.

Comparison of color

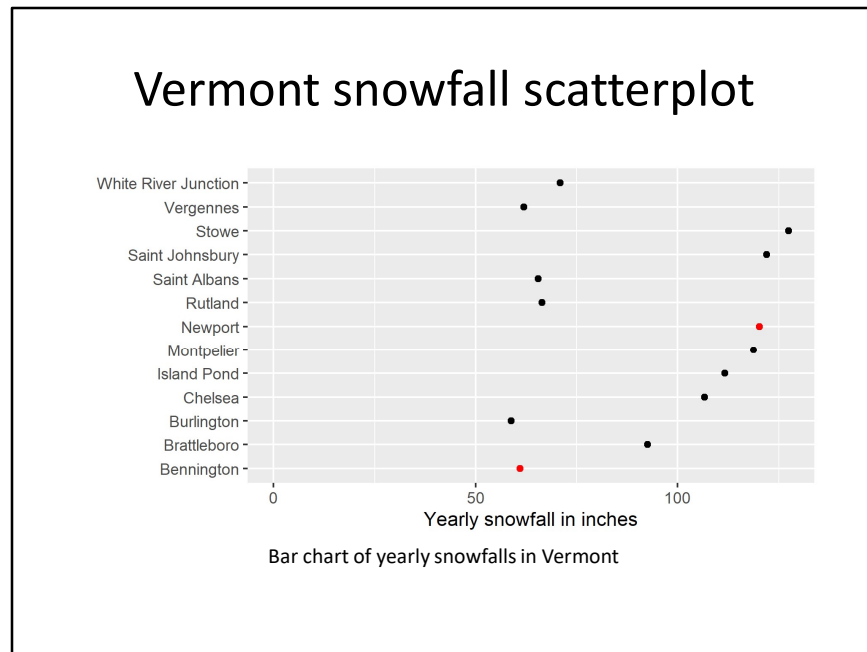


Okay now, which city has a higher average yearly snowfall?

How much more?

Well the blue surrounding St. Albans looks to be a bit darker than the blue surrounding Bennington. So maybe it has more snow.

Now that wasn't too hard, was it?



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

But notice how much easier that question becomes when you display the snowfalls as scatterplots?

I've highlighted Bennington and St. Albans to make it a bit easier for you, but the answer is almost immediate because you are determining relative position versus shades of a color.

Now there are some caveats here. The scatterplot doesn't help much if you live outside the twelve cities listed here. You also lose the ability to see a ridge of higher snowfall running right down the middle of the state. There's also a huge amount of snow in the parts of New York just east of Lake Ontario.

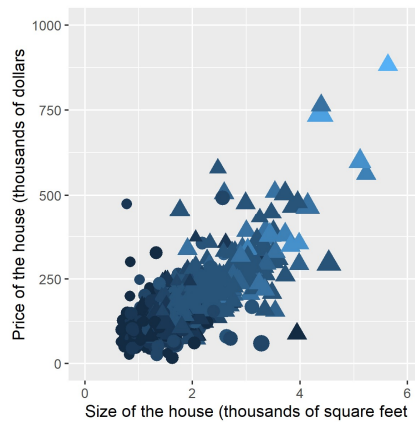
So don't take this too literally, but in general some features of a graph lend themselves better to quicker and more accurate answers.

Break #1

- What have you learned so far
 - Two quantifiable measures: speed and accuracy
 - Hierarchy of perception
 - Relative position is easy to judge
 - Difference in color are harder to judge
- What's coming next
 - The five dimensions of a scatterplot

Let's take a break here and see if there are any questions.

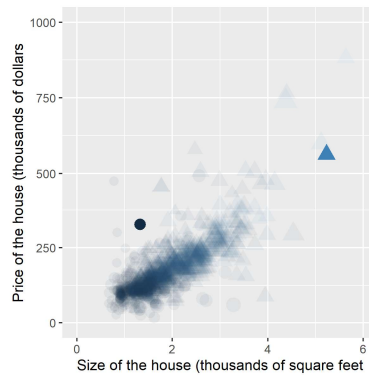
A five dimensional scatterplot



Here's a plot that shows the five features that you have available in a scatterplot. This is not a plot format that I would recommend, just one that illustrates a basic concept.

The dataset comes from a website called DASL which is short for Data And Story Library. It represents a survey of housing prices in Saratoga, New York. For each house sale, there is information about the house, such as square footage, number of bedrooms, number of bathrooms, and whether the house has a fireplace.

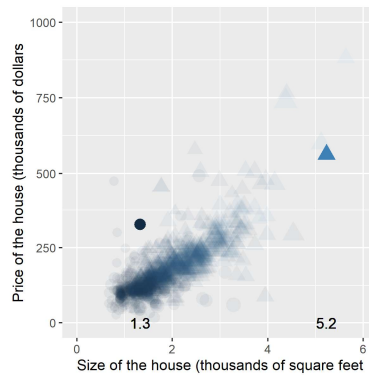
Highlight two marks



Plot of snowfall in Vermont with two marks highlighted

I want to highlight just two data marks in this graph to show what the five dimensions are going to be.

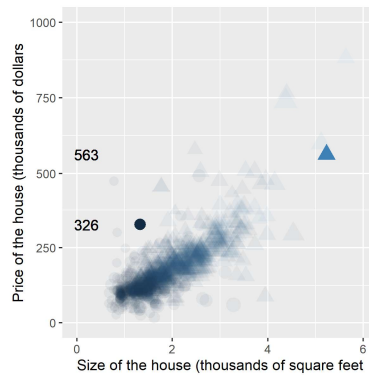
X dimension <- Square footage



Plot of snowfall in Vermont emphasizing X dimension

The X dimension shows the size in thousands of square feet for each house. The two houses highlighted have 1,300 and 5,200 square feet.

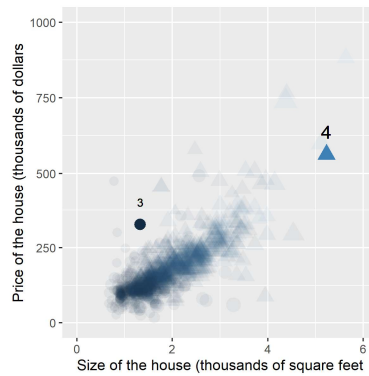
Y dimension <- Price



Plot of snowfall in Vermont emphasizing Y dimension

The Y dimension shows the price in thousands of dollars for each house. The two houses highlighted have 1,300 and 5,200 square feet.

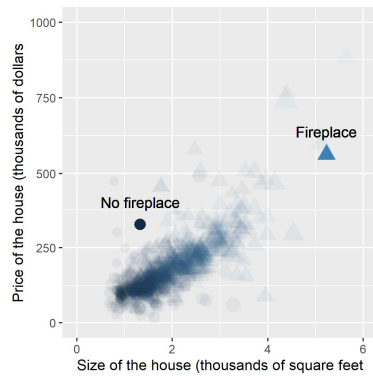
Size dimension <- Number of
bedrooms



Plot of snowfall in Vermont emphasizing size dimension

The size of the data mark is proportional to the number of bedrooms. The smaller of the two highlighted marks is a house with three bedrooms. The larger is a house with four bedrooms.

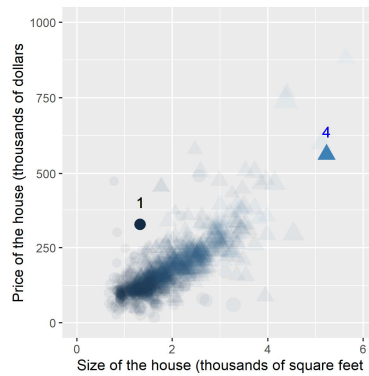
Shape <- Fireplace indicator



Plot of snowfall in Vermont emphasizing shape dimension

The shape of the mark indicates whether the house has a fireplace. A triangle represents a house with a fireplace and a circle represents a house with one without a fireplace.

Color <- Number of bathrooms



Plot of snowfall in Vermont emphasizing color dimension

The color of the data mark indicates how many bathrooms a house has. The black data mark indicates only one bathroom and the blue data mark indicates four bathrooms.

Break #2

- What have you learned
 - Five dimensions of a scatterplot (x, y, size, shape, and color)
- What's coming up next
 - Mitigating the problem of overprinting

Let me stop again and see if there are any questions.

X and Y position, mitigating overprinting

– Partial solutions

- Open symbols
- Small size
- Log transformation
- Opacity
- Jittering

The biggest problem with many graphs is overprinting. So much data ends up producing a big black uninterpretable blob. There are several solutions that can help somewhat with overprinting.

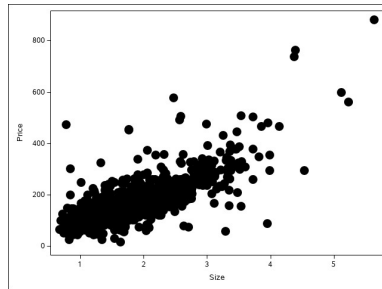
Mitigating overprinting: use open symbols



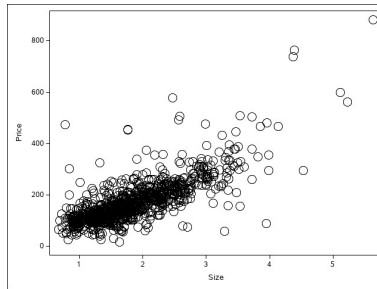
Comparison of solid versus open circles

If you look at the figure to the left, it looks like it might be three or maybe four data marks all clustered together. There is a bit of an indentation in the southeast corner of this blog that gives a hint that it is really four rather than three data marks. The figure on the right uses open circles and you can tell much faster that there are indeed four data marks.

Mitigating overprinting: use open symbols



Original scatterplot



Scatterplot with open circles

Here's what open circles do to a scatter plot of square footage versus price. It only helps a little, but you can start to see the difference between regions that have a moderate amount of data versus a much larger amount of data.

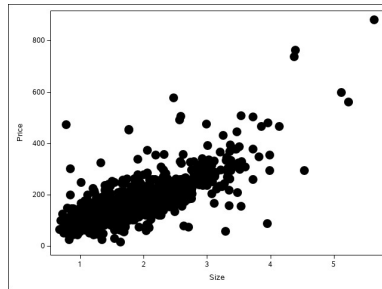
Mitigating overprinting: Use small size marks



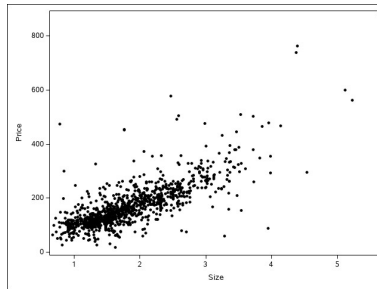
A cluster of four data marks comparing larger versus small

If you use smaller data marks, you will be able to separate out the four individual data marks in this cluster. There are disadvantages to smaller data marks. They sometimes make it easier for you to ignore outliers. They also cause problems when you want to use different shapes, as the shapes become less distinguishable for smaller sizes. But they work well for mitigating overprinting.

Mitigating overprinting: use smaller size marks



Original scatterplot



Scatterplot with smaller sized symbols

In the square footage by price scatterplot, the smaller marks make it much easier to see where the individual data points are.

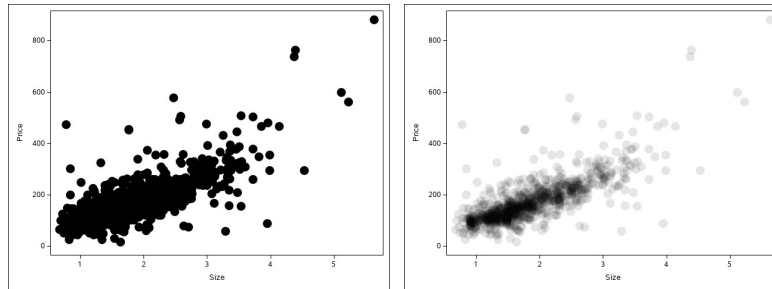
You can shrink this down to a single pixel, if you like. The plot looks like dust specks. Sometimes that ends up revealing a lot more of what is going on near the center of the data blob.

Mitigating overprinting: Opacity



A cluster of four data marks varying the opacity

Mitigating overprinting: Opacity

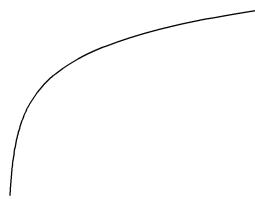


Original scatterplot

Scatterplot with opacity

Here is a comparison to fully opaque data marks to semi-transparent data marks. The effect can sometimes minimize the effect of outliers, but it does help reveal the structure of the

Mitigating overprinting: Log scale

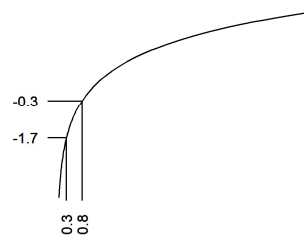


Log function

Often your data is crowded in the lower left corner of your graph. This is caused by skewness in your variables. A log transformation can often help in such a situation.

Here's a picture of the log function. It is steep on the left and closer to flat on the right.

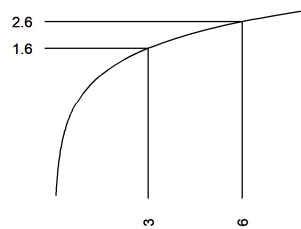
Mitigating overprinting: Log scale



Log transformation of small values

The steepness on the left means that the log function tends to stretch out small values. The data points 0.3 and 0.8 are half a unit apart, but after the log transformation they are 1.4 units apart.

Mitigating overprinting: Log scale



Log transformation of large values

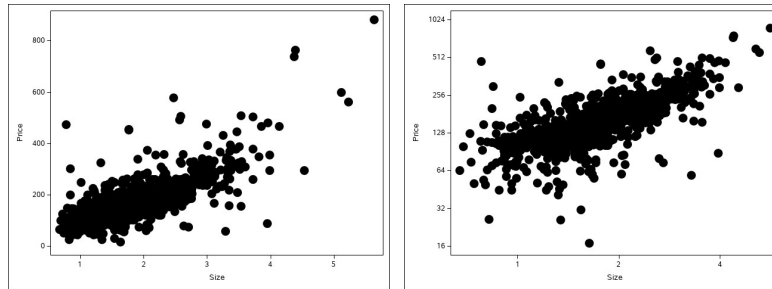
The relative flatness on the right means that the log function tends to squeeze large values together. The values of 3 and 6 are 3 units apart, but after a log transformation, they are only 1 unit apart.

By stretching the values stuck in a large blob in the lower left corner of the graph and squeezing the few outlying values elsewhere, you often end up with a spread of data that is much more uniform across the plotting area.

A small hint: don't try a log transformation, unless your data has a large relative range. If your largest value is not at least three times as large as your smallest value, then the log transformation is unlikely to have an impact.

Also, you can't use a log transformation if you have zeros or negative values in your data.

Mitigating overprinting: Log scale

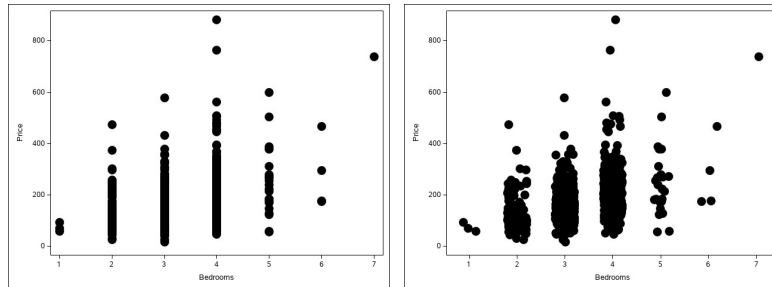


Original scatterplot

Scatterplot with log scale

Notice how the plot on the log scale tends to fill up the plotting area a bit more efficiently. This helps reduce overprinting a bit.

Mitigating overprinting: Jittering



Plot of bedrooms and price without jittering

Scatterplot with jittering

I have to switch to a different set of variables to better illustrate jittering. The Y axis is still price, but the X axis is now the number of bedrooms, and most of the data values are now squished together into vertical lines.

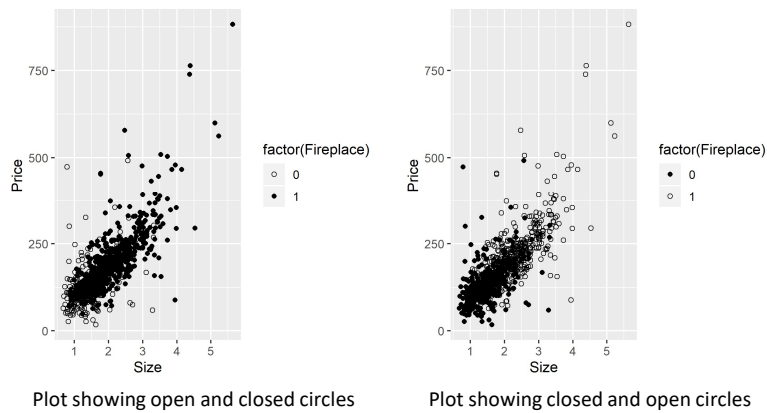
You could use a boxplot here, but jittering is also effective. Jittering is a random shift of datapoints to the left or right (sometimes up and down as well) to help spread the data out a little bit.

Break #3

- What you have learned
 - Strategies for mitigating overprinting
- What's coming next
 - Issues with shape and size

Let's stop and see if there are any questions.

Problem with open and closed shapes together



In a fight between open and closed symbols, the closed symbols always win. Notice in the left side plot, it looks like most of the houses have fireplaces (Fireplace=1 is the closed circle). In the second plot, it looks like most of the houses do not have fireplaces (Fireplace=0 is the closed circle). In fact the split is pretty close to even. 60% of the houses have fireplaces.

Don't use too many shapes



Plot with seven different shapes

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.

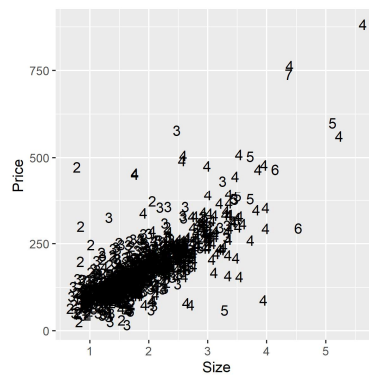
An alternative with only two shapes



Plot with only two shapes

Here is a better plot. You lose a bit of information by using only two shapes, but the plot is simpler and easier to interpret.

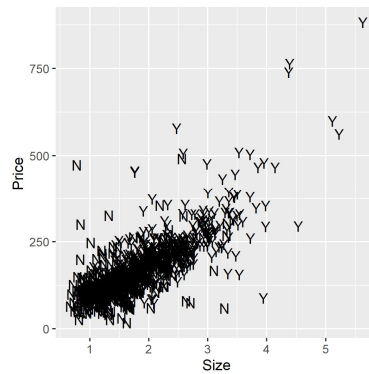
Using text instead of symbols



Plot using numbers in place of shapes

Sometimes a simple bit of text works better than symbols. In most software system, it is pretty easy to print the actual number where mark should be. It works great for single digit numbers, and sometimes even for two digit numbers

Using text instead of symbols

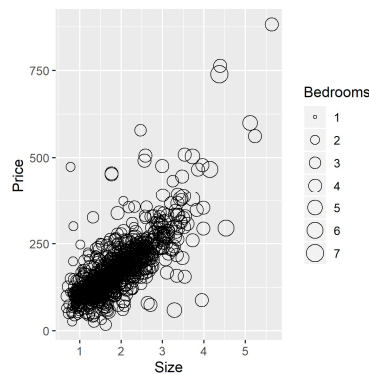


Plot using letter codes Y for yes and N for no

Here's another example using letters.

Other examples might be M for male and F for female, T for treatment and C for control.

Example using size: the bubble chart



Plot with size representing number of bedrooms

This plot shows how the size of the mark can indicate a third variable. In this case, it is the number of bedrooms. This type of plot is often called a bubble chart. We've seen far too many bubble charts since the start of the COVID-19 crisis, and those bubbles are getting bigger all the time.

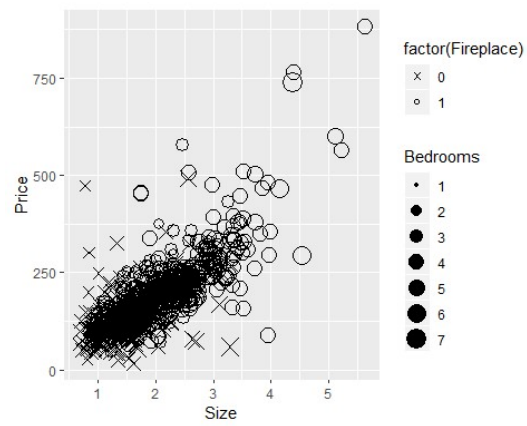
Size

- Never for a categorical variable
- Proportional to
 - Diameter?
 - Area?

Never use size for a categorical variable. You want all the categories to be equally distinguishable, and categories given the middle sizes end up looking too much alike.

If size represents a continuous variable, do you want the size to be proportion to the diameter of your circle (or the length of your square)? Or would it be better to make the area of your circle or square proportional to the data value. There's been a lot written on this, and the general consensus is to use area.

Size and shape don't mix

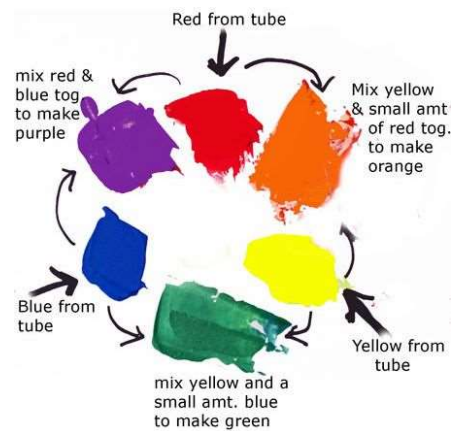


Break #4

- What have you learned
 - Issues with shapes and sizes
- What's coming next
 - Issues with color

Let's stop again and see if there are any questions.

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



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 hexadecimal of all zeros is pure black and at the other extreme, a hexadecimal of all F's is pure white.

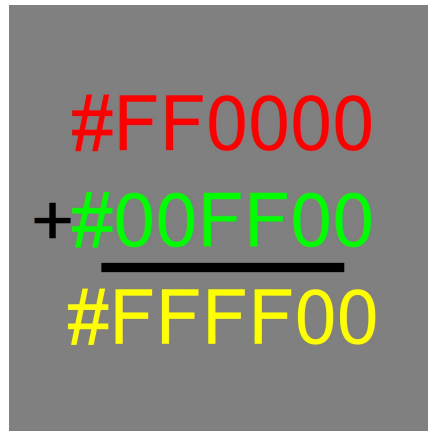
The first two hexadecimal digits represent the red channel. The highest value FF for the red channel combined with zeros for the other two channels (#FF0000) equals pure red.

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

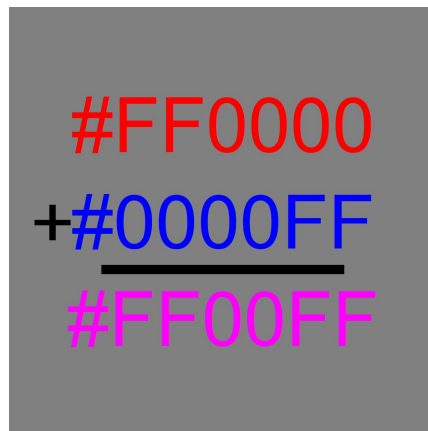


A diagram illustrating the addition of red and green colors to produce yellow. It features a gray rectangular background. On this background, the red hex code `#FF0000` is written in red text. Below it, a plus sign `+` is followed by the green hex code `#00FF00` in green text. A horizontal line is drawn under the green hex code. Below the line, the resulting yellow hex code `#FFFF00` is written in yellow text.

$$\begin{array}{r} \text{\#FF0000} \\ + \text{\#00FF00} \\ \hline \text{\#FFFF00} \end{array}$$

When 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 channel), you get yellow, which is FF in both the red and green channels.

Colors, Red plus blue equals
magenta

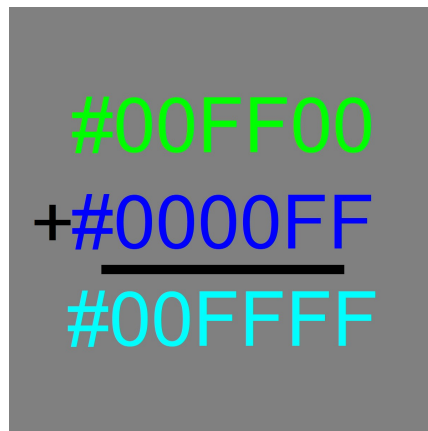


A diagram illustrating the addition of red and blue to create magenta. It features a gray rectangular background. On this background, the red hex code `#FF0000` is written in red text. Below it, a plus sign `+` is followed by the blue hex code `#0000FF` in blue text. A horizontal line is drawn under the blue hex code. Below the line, the resulting magenta hex code `#FF00FF` is written in magenta text.

$$\begin{array}{r} \text{\#FF0000} \\ + \text{\#0000FF} \\ \hline \text{\#FF00FF} \end{array}$$

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

Colors, Green plus blue equals cyan

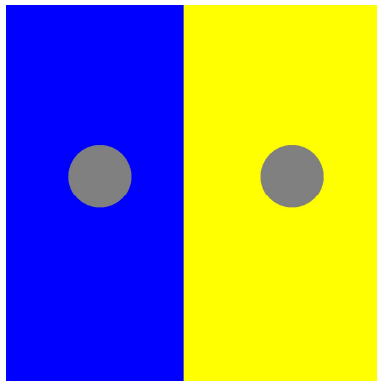


A diagram illustrating the addition of green and blue to produce cyan. It features a dark gray rectangular background. On this background, the text is arranged vertically: the green hex code `#00FF00` is at the top in green; a plus sign `+` followed by the blue hex code `#0000FF` is in the middle in blue, with a short horizontal black line underneath the blue code; and the resulting cyan hex code `#00FFFF` is at the bottom in cyan.

$$\begin{array}{r} \text{\#00FF00} \\ + \text{\#0000FF} \\ \hline \text{\#00FFFF} \end{array}$$

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

Basic colors have different luminance

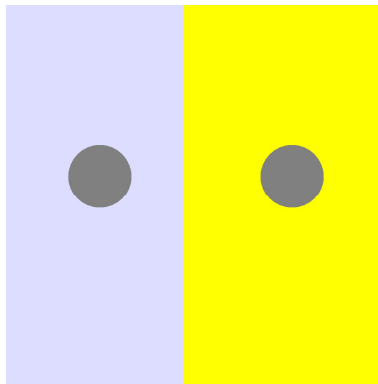


Gray dot on blue and yellow backgrounds

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.

An attempt to equalize luminance



Using a lighter shade of blue to equalize luminance

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 background 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 fashion 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.

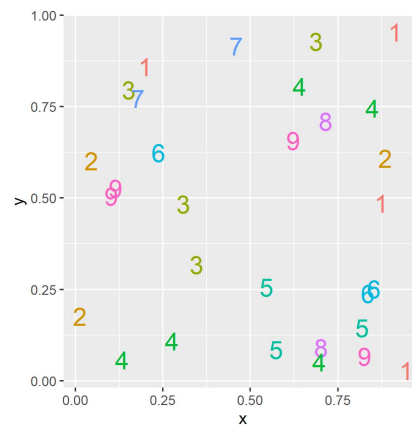
Text with a variety of colors

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

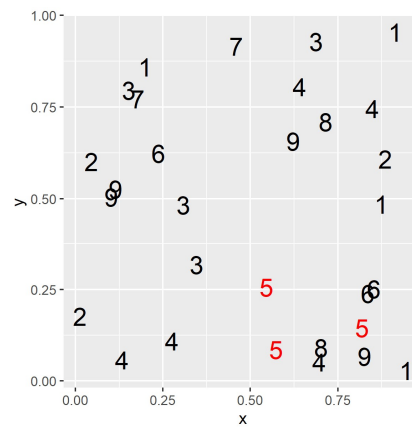


Here's an exercise that adapted from Olson and Bergen.

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.

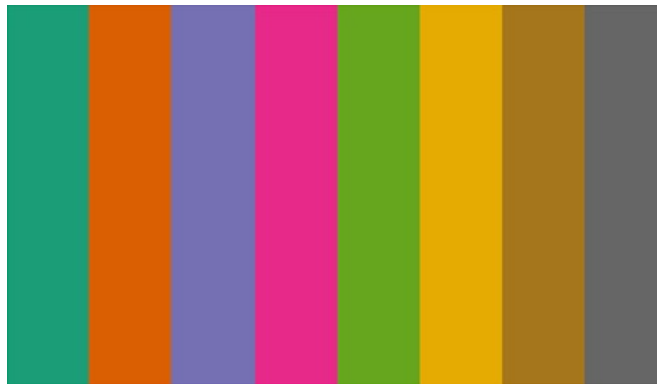
Did the colors help? Well, not all that much. It is hard to pick out nine colors and not have a few of them look very similar. In particular, the 5's and the 6's are pretty close, as are the 8's and the 9's.

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.

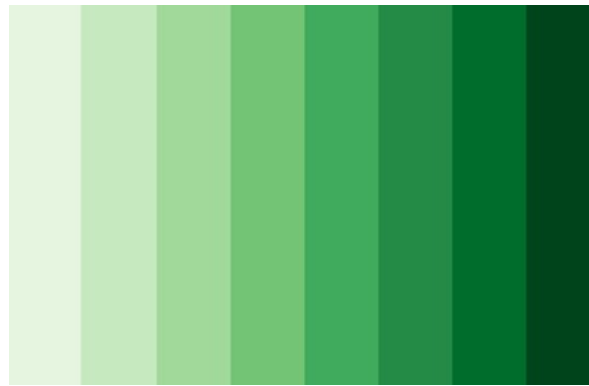
Discrete colors



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

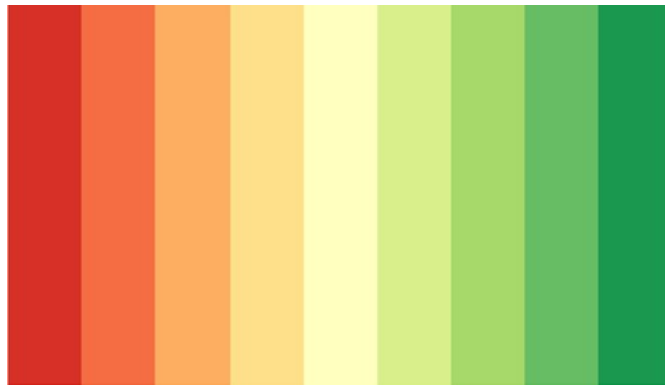


Simple gradient from RColorBrewer

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

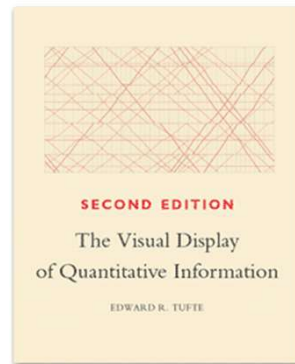
A diverging gradient moves between two distinct colors and takes a side trip in the middle to a third color. Typically, the middle color in a diverging gradient has a much higher luminance or a much lower luminance than the two extremes and is intended to fade into the background. The diverging gradient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

Break #5

- What have you learned
 - Issues with color
- What's next
 - Recommended books

Just one more break. Are there any questions?

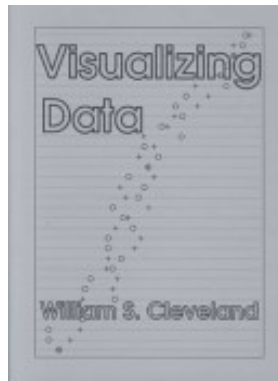
Recommended books



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.

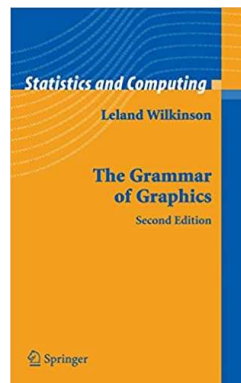
Recommended books



Book cover of Visualizing Data

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.

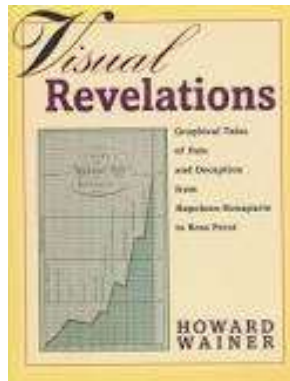
Recommended books



Book cover of Grammar of Graphics

This book has had a tremendous impact on data visualization software, but it took ten years because no one really understood what the book was trying to say. It is a very difficult read, full of theory. It is probably best for those who write software rather than those who use it.

Recommended books



Book cover of Visual Revelations

Howard Wainer is a great story teller and you learn a lot from each story that he tells. This book takes a look at data visualization from many many years ago and points out lessons that you can learn from this. It's sort of like teaching through analogy.

Recommended books



Book cover of Creating More Effective Graphs

This book is very applied and is an excellent starter book for anyone wanting to learn more about data visualization. I found it a bit simplistic in parts, but after having read so many other books, perhaps that is unavoidable.

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.