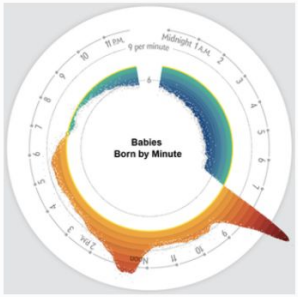
Data Visualization for Scientific Discovery

zan.armstrong@gmail.com
@zanstrong

I'm a [data visualization designer and engineer](#).

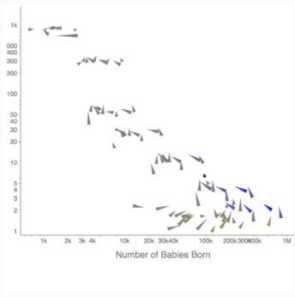
ZAN ARMSTRONG

WORK WRITING/SPEAKING RESOURCES FOR TEACHERS ABOUT

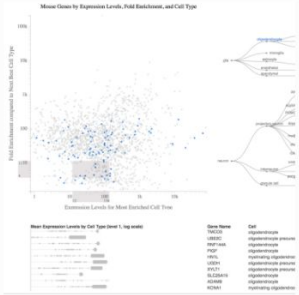


SCIENTIFIC AMERICAN: WHY ARE SO MANY BABIES BORN AROUND 8AM?

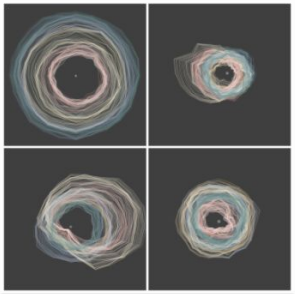
WITH THE TEAM AT STAMEN



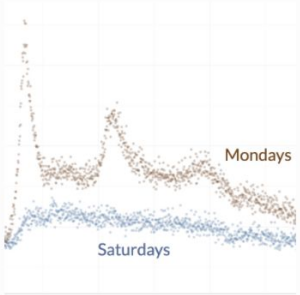
RESEARCH: VISUALIZING STATISTICAL MIX EFFECTS AND SIMPSON'S PARADOX



GENETIC EXPRESSION TOOL



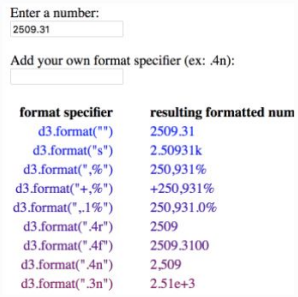
WEATHER CIRCLES



OPENVISCONF: EVERYTHING IS SEASONAL

WHICH IS BIGGER?

SCHOOL FOR POETIC COMPUTATION



TINY TOOLS

I'm a data visualization designer and engineer.

I am part of **Google's Accelerated Science** team.

My goal is to use data visualization to help a team of scientists, statisticians/data scientists, and machine learning engineers analyze their data and make discoveries.

All you need to know about Data Viz

Pie Charts are bad.

Non-zero bar charts are bad.

Dual y-axis are bad.

Rainbows are bad.

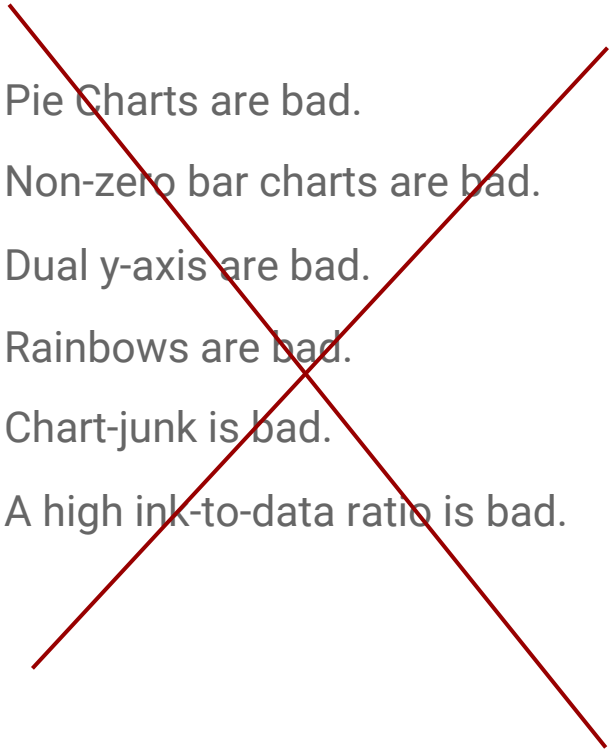
Chart-junk is bad.

A high ink-to-data ratio is bad.

The End.

Any questions?

Oft-quoted "rules" of data viz.




CALORIES
ACRYLAMIDE
FAT
SALT
PRESERVATIVES
CARBOHYDRATES

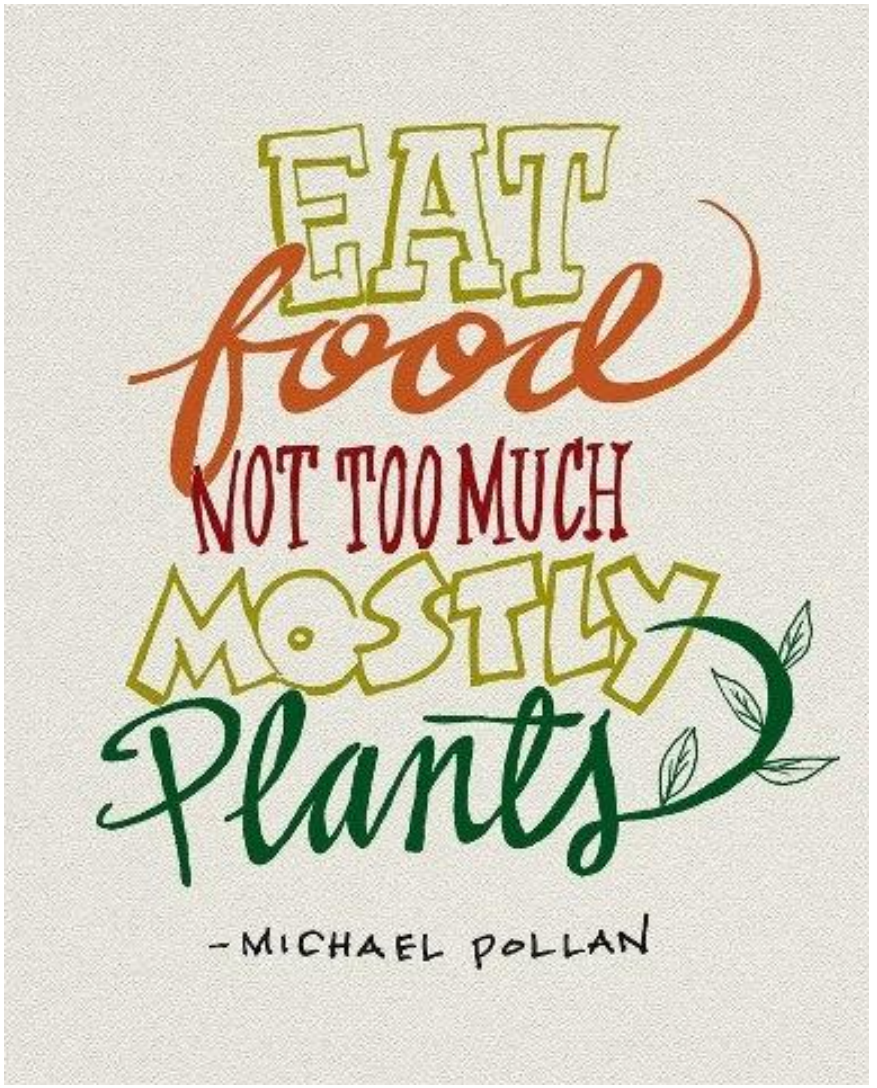

is SOY
BAD
for YOU?


ARE
PRESERVATIVES
BAD FOR
YOU?


How Food Coloring Is Made
& Is It Bad For You?

Better to have a few good guidelines than a bunch of rules telling you what not to do

Better to have a few good guidelines



Better to have a few good guidelines

Create small multiples (many, related charts)

Use color intentionally

Optimize your charts for your time/energy/attention

Is it good or bad depends on your goals, context, and constraints.

The answer to "is it good" depends on goals, context, and constraints.

Is a candy bar good for me?

The answer to "is it good" depends on goals, context, and constraints.

Is a candy bar good for me?

Am I at home drinking tea?

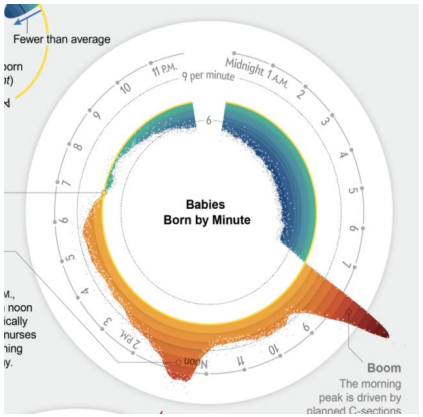
no :(

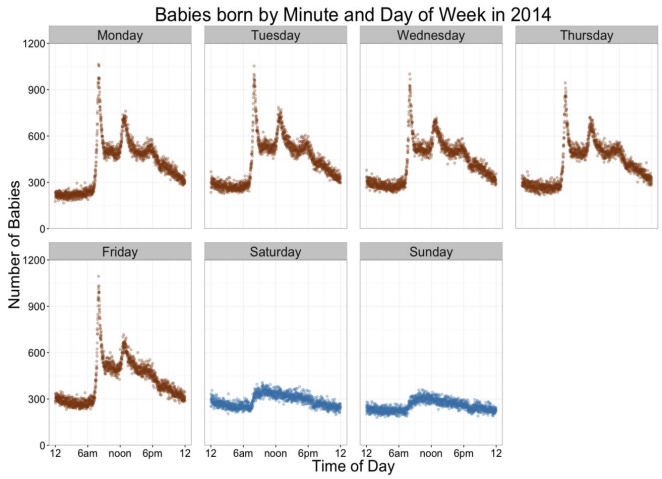
Am I on a week-long backpacking trip?

yes!

The answer to "is it good" depends on goals, context, and constraints.

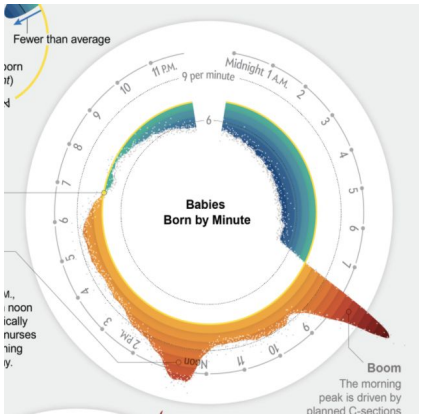
Which visualization is better?

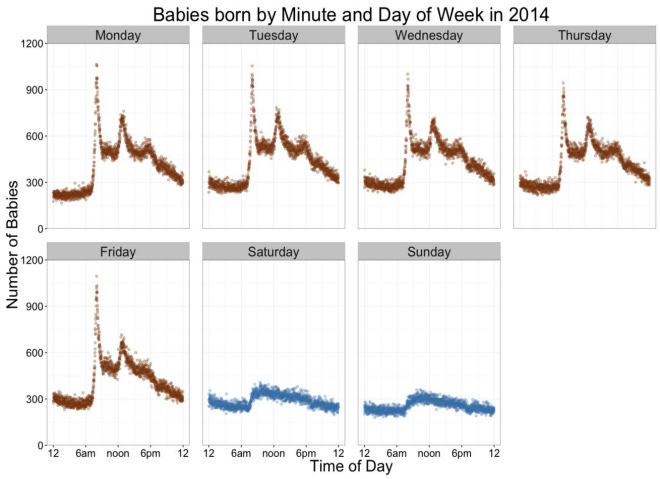




The answer to "is it good" depends on goals, context, and constraints.

Which visualization is better?





It depends!

What's the point?

How much time to create?

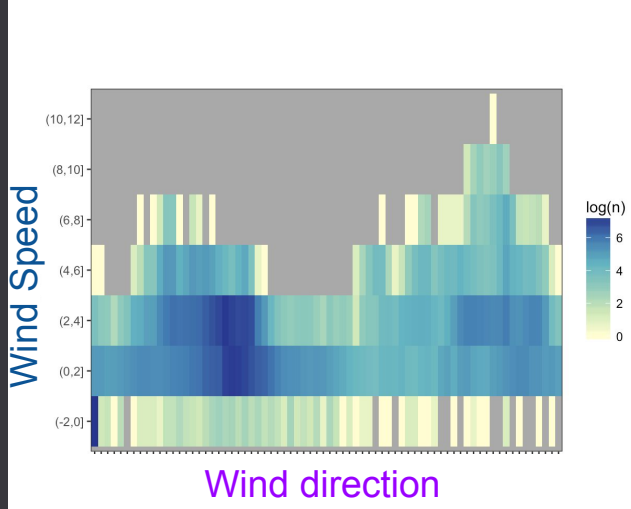
Audience? Who? Their goal?

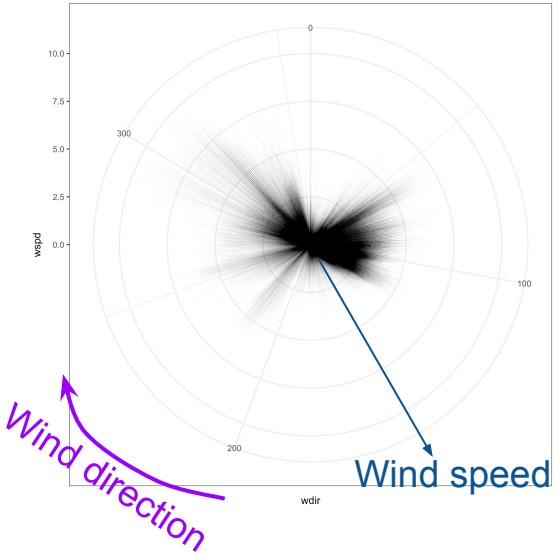
Do I care about day of week?

Do I care about total values or difference to the mean?

The answer to "is it good" depends on goals, context, and constraints.

Which visualization is better?





The answer to "is it good" depends on goals, context, and constraints.

For communication (publication, journalism, etc)

For scientific analysis/discovery (exploratory & "pipelines")

The answer to "is it good" depends on goals, context, and constraints.

For communication (publication, journalism, etc)

Does it attract the audience's attention?

Get the message across?

It is true to the data? (not misleading)

How much space on the page or screen?

Is it immediately understandable?

Can everybody interpret it?

Most viz "rules" optimized for communication

For scientific analysis/discovery (exploratory & "pipelines")

The answer to "is it good" depends on goals, context, and constraints.

For communication (publication, journalism, etc)

Does it attract the audience's attention?

Get the message across?

It is true to the data? (not misleading)

How much space on the page or screen?

Is it immediately understandable?

Can everybody interpret it?

Most viz "rules" optimized for communication

For scientific analysis/discovery (exploratory & "pipelines")

How much of your time/energy is needed to interpret?

If something in your data is important, will you see it in the viz?

How long does it take to create?

How much mental effort does it take to create?

How hard is it to ask a new question?

Will I look at this type of chart/tool often? Aka - is a learning curve for interpretation ok?

Goals, context, and constraints for analysis and discovery

If something in your data is important, will you notice it?

It's ok to risk something looking important that isn't, because you're going to do more analysis.

Goals, context, and constraints for analysis and discovery

Your time & energy is your most limited resource.

Number of charts is (generally) not a limiting resource.

Is it good or bad depends on your goals, context, and constraints

Is it good or bad depends on your goals, context, and constraints*

*for food and viz... and engineering, ML, and stats

Create small multiples

What are small multiples?

What are small multiples?

Lots of small charts.

What are small multiples?

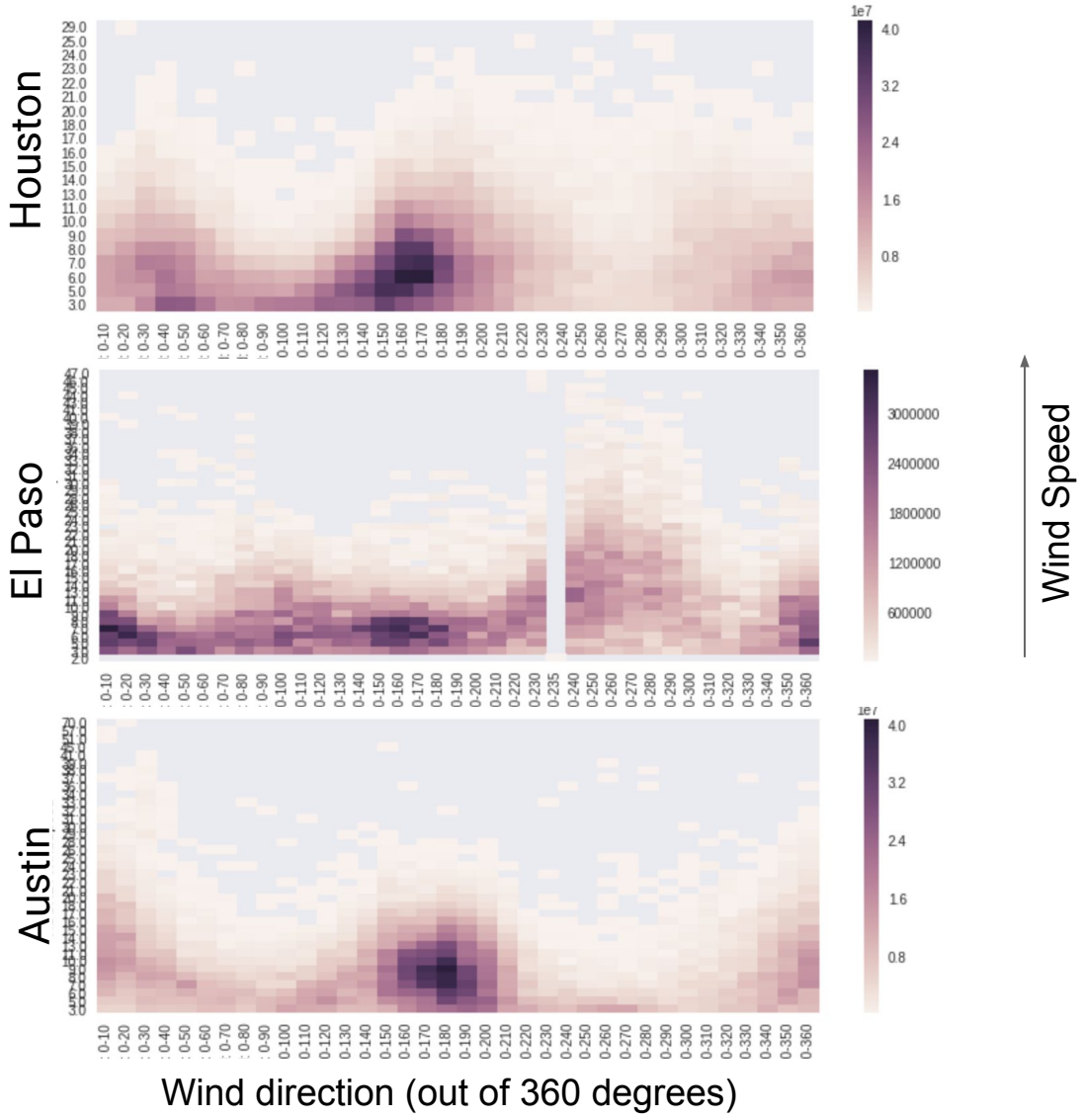
Lots of small charts.
Similar to each other.

What are small multiples?

Lots of small charts.
Similar to each other.
That you look at at the same time.

What are small multiples?

Lots of small charts.
Similar to each other.
That you look at at the same time.



Why small multiples?

You can't compare things you can't see.

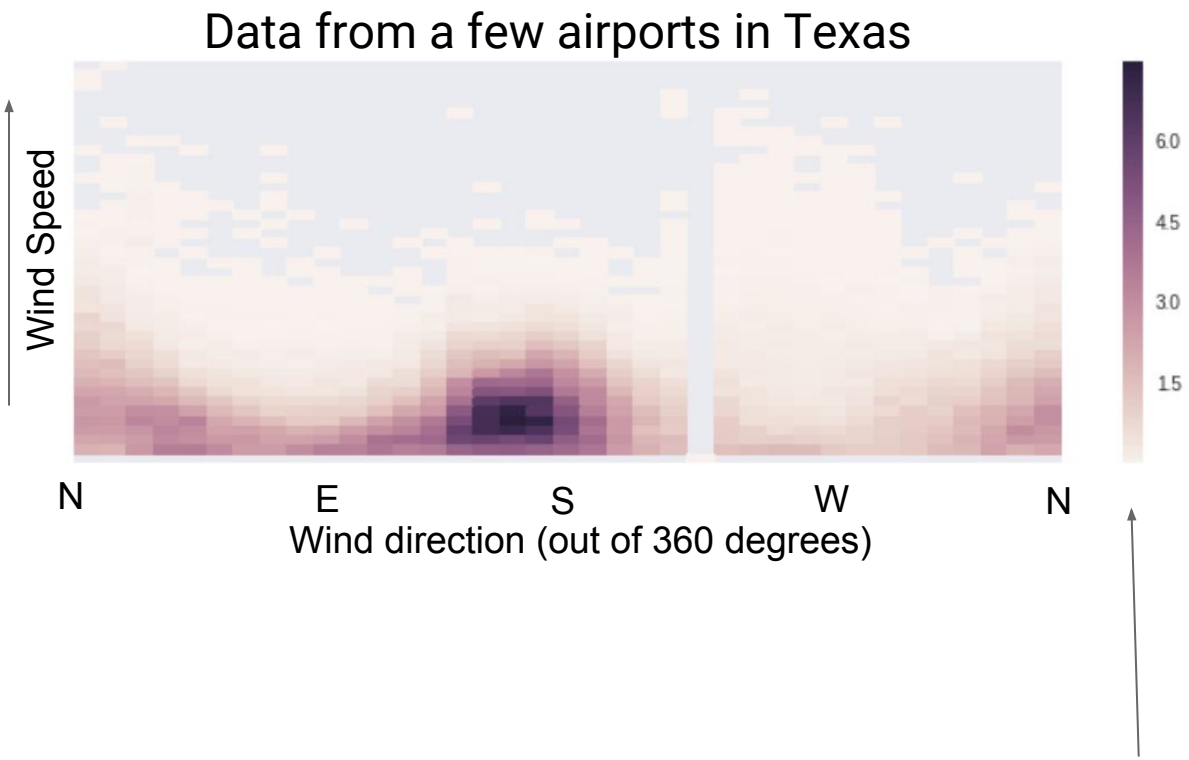
Create small multiples

Same form, related data

Small Multiples

Same form, Related Data

Let's say we're interested in the relationship between wind direction and wind speed in Texas.

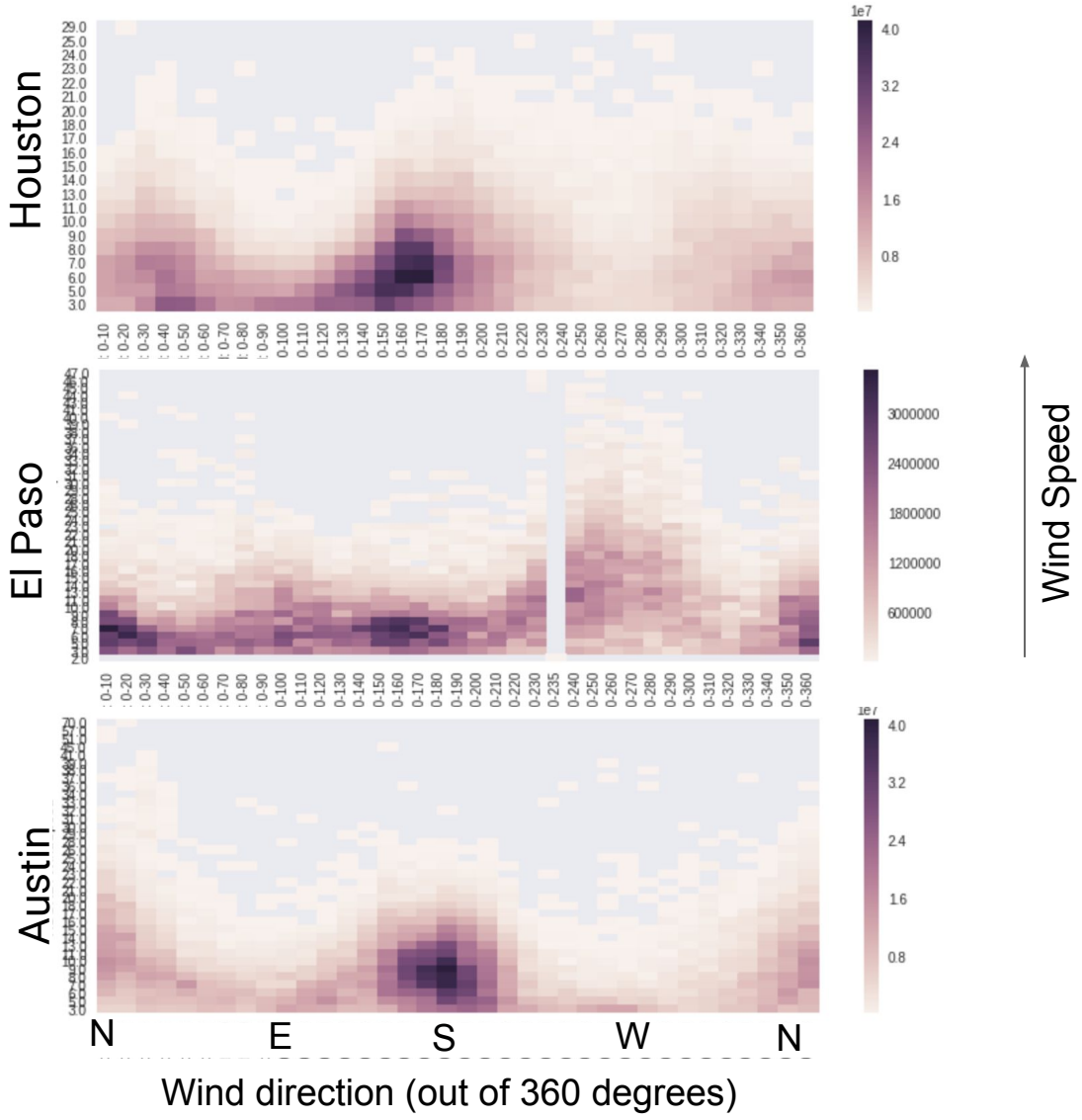


# of times that wind speed/wind direction has been observed in the past 3 years.

Small Multiples

Same form,
Related Data

Small multiples provide **context** and **contrast**.



Create small multiples

Same data, different form

Small Multiples

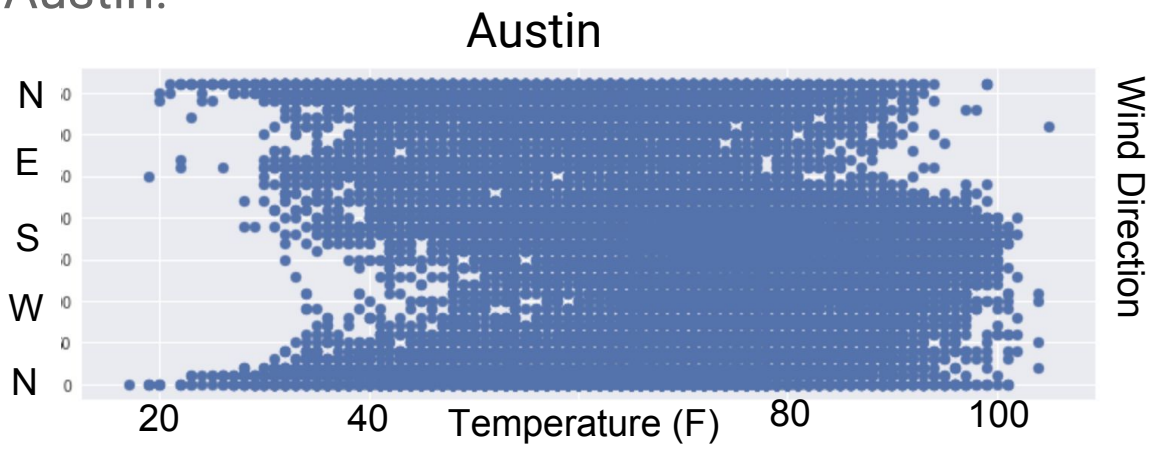
Same data, Different form

Let's say we're interested in the relationship between wind direction and temperature in Austin.

Small Multiples

Same data,
Different form

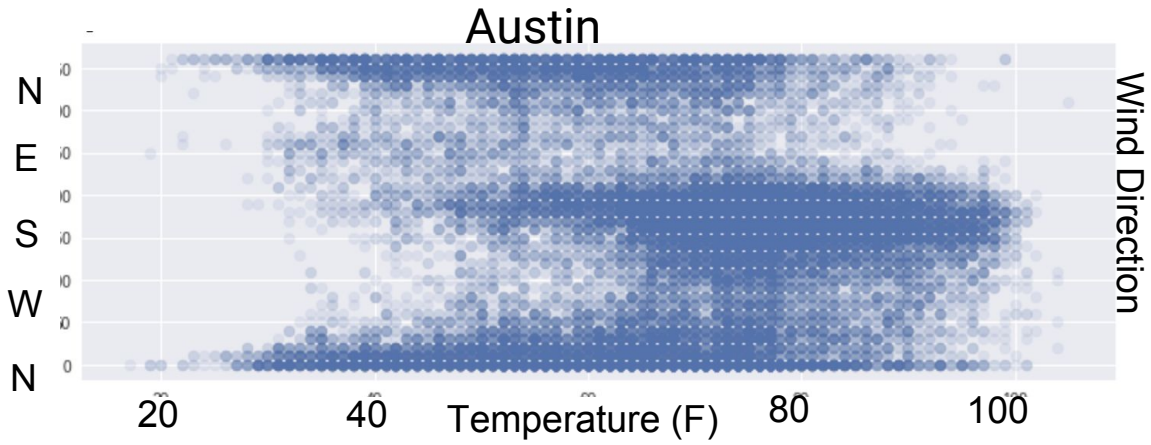
Let's say we're interested in the relationship between wind direction and temperature in Austin.



Small Multiples

Same data,
Different form

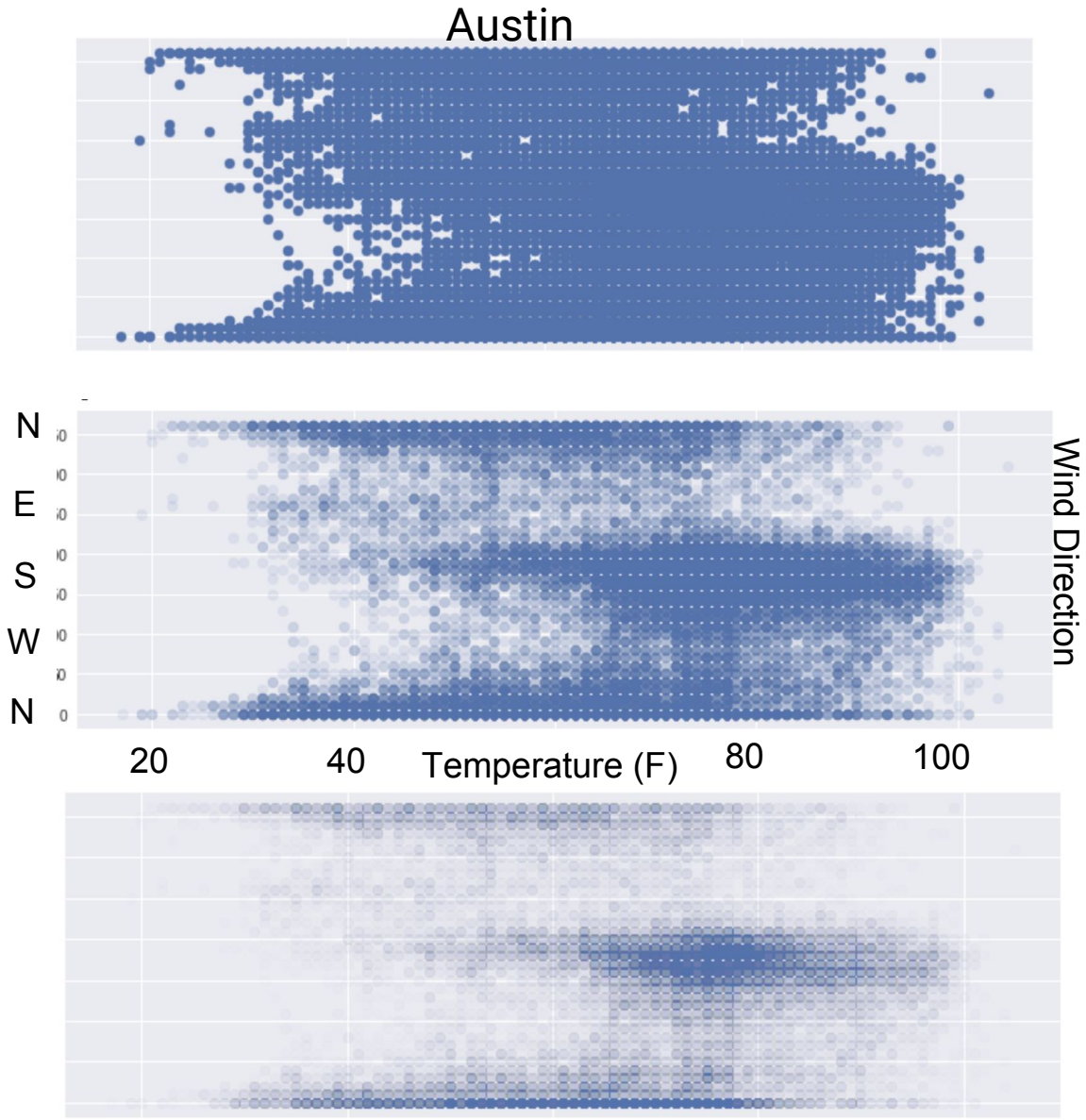
Common trick: Lower opacity to see density



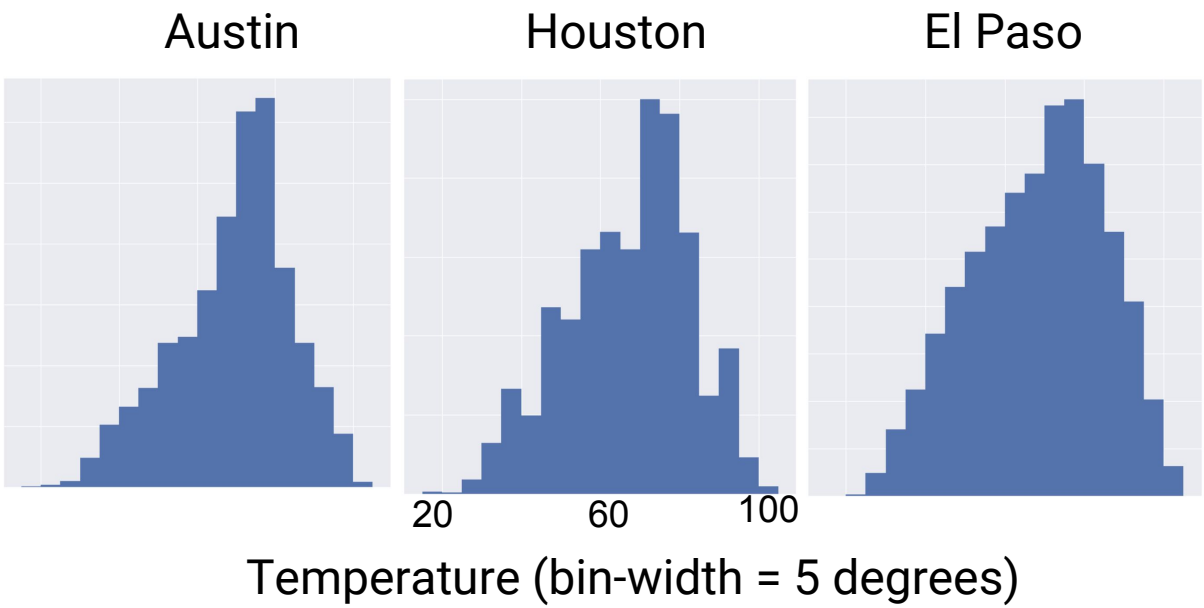
Small Multiples

Same data,
Different form

Let's look at 3 versions at once: see extremes, and highest density of data.

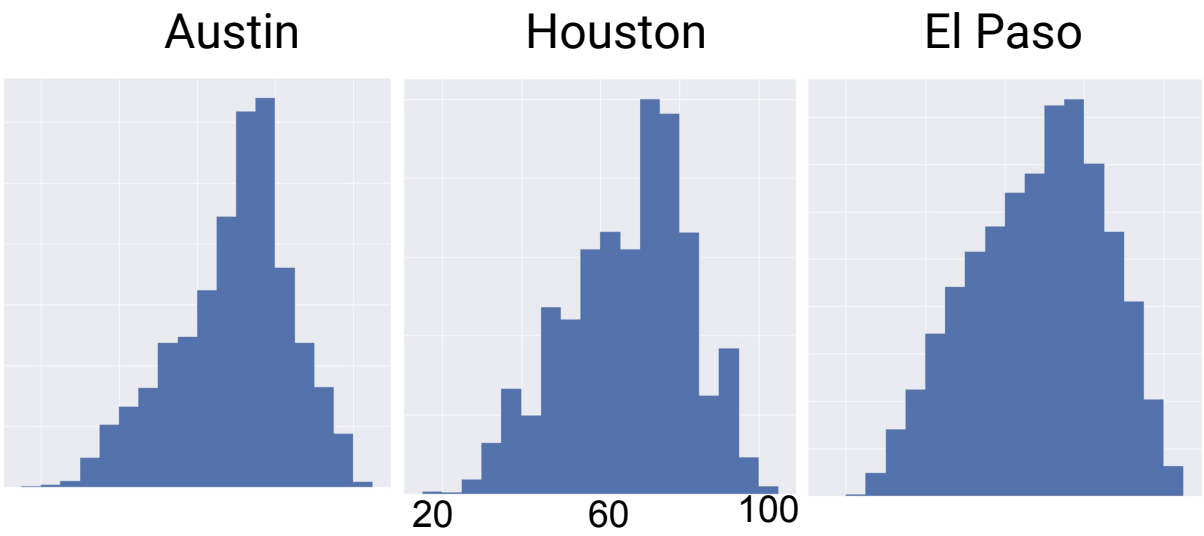


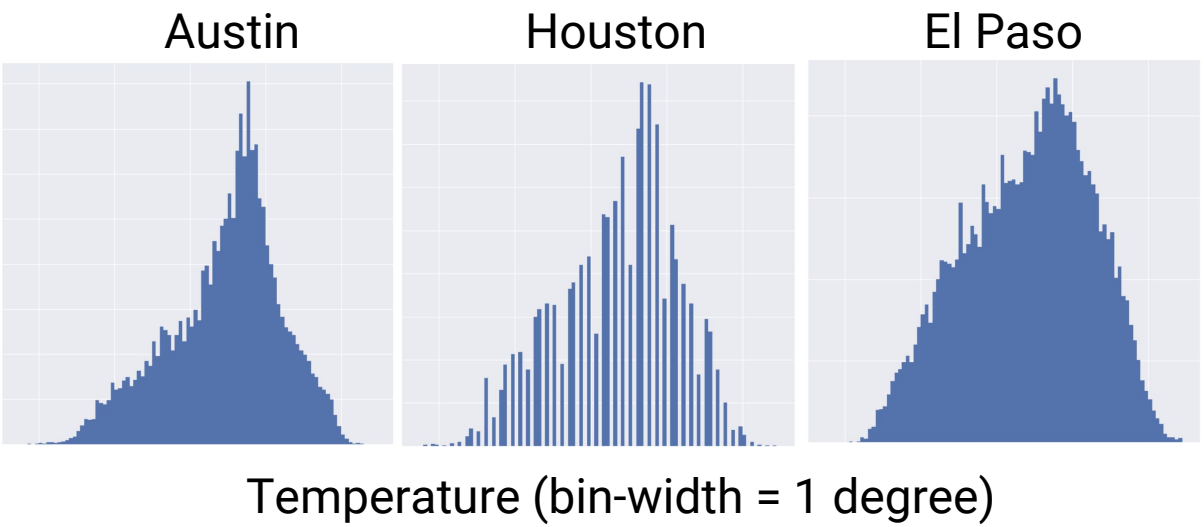
Small Multiples



For more, see Amelia and Arun's exploration of binwidth and histograms

Small Multiples





For more, see Amelia and Arun's exploration of binwidth and histograms

Small Multiples

Resources

Matplotlib: Subplots Example (most flexible)

Seaborn: FacetGrid

Altair: Trellis Example

Use color intentionally

The basic visual attributes … are perceived without any conscious effort.

[This] is called "preattentive" processing…

Preattentive perception is done in parallel, but attentive processing is done serially and is, therefore, much slower.

Stephen Few - Perceptual Edge

The basic visual attributes … are perceived without any conscious ***effort****.*

[This] is called "preattentive" processing…

Preattentive perception is done in parallel, but attentive processing is done serially and is, therefore, much slower.

Stephen Few - Perceptual Edge

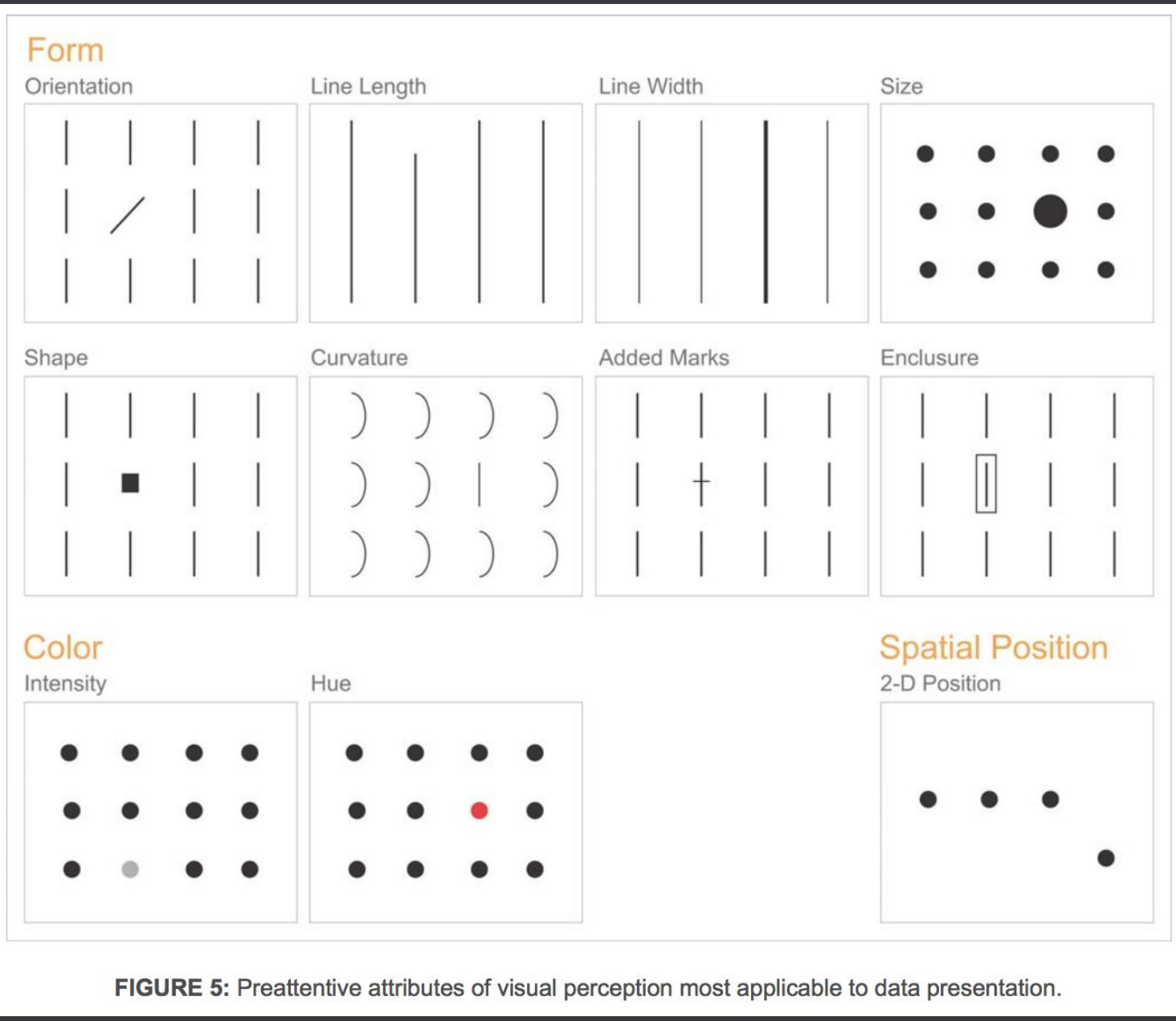


FIGURE 5: Preattentive attributes of visual perception most applicable to data presentation.

Stephen Few - Perceptual Edge

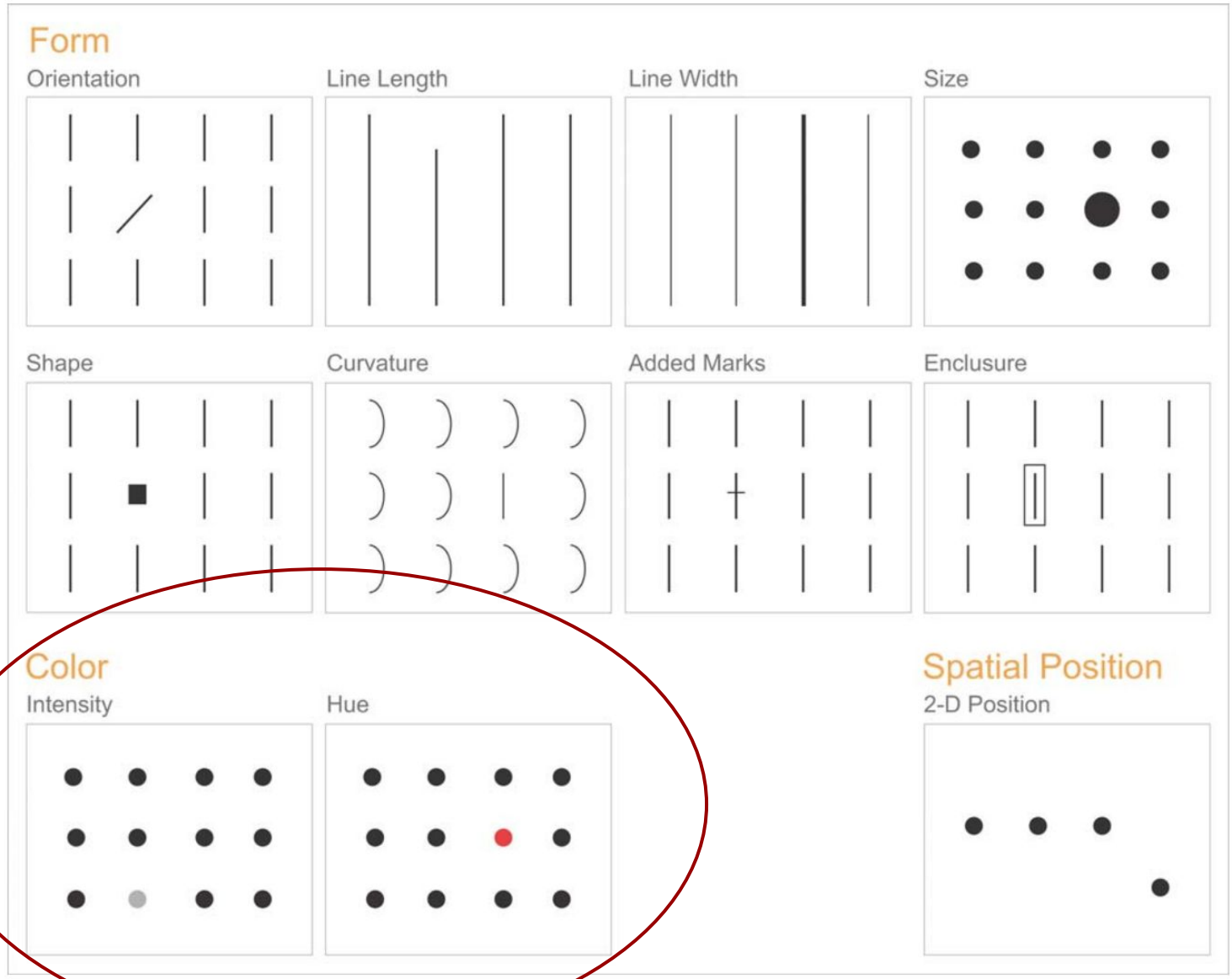


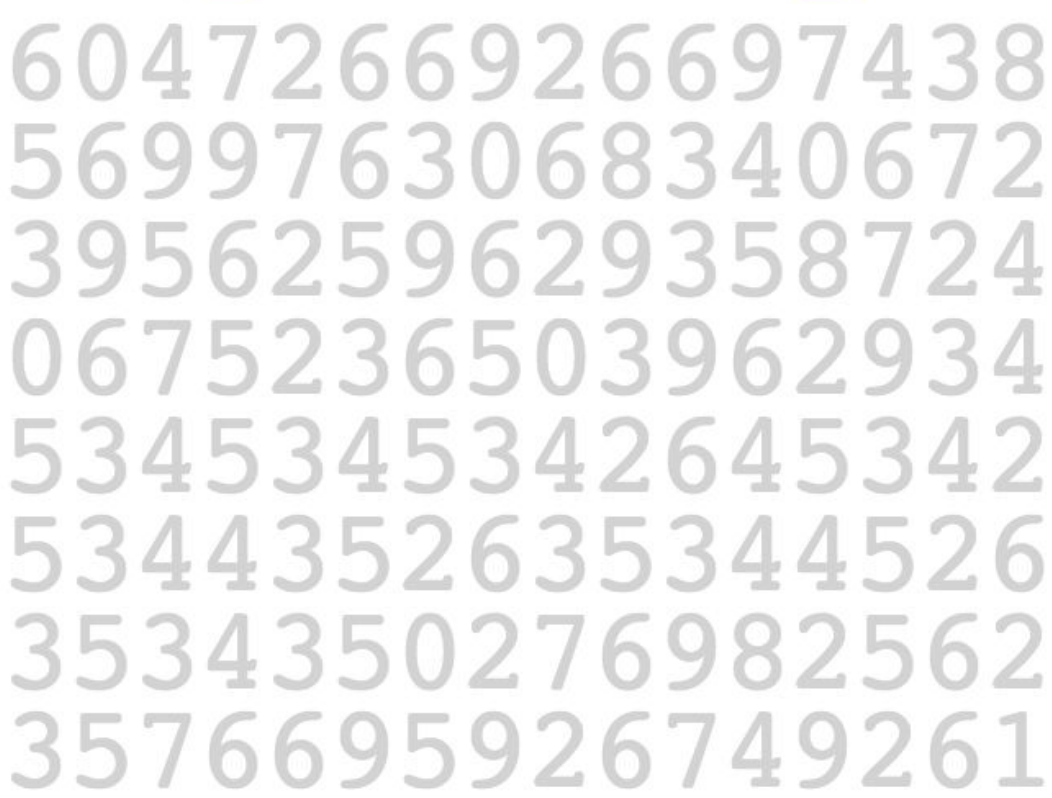
FIGURE 5: Preattentive attributes of visual perception most applicable to data presentation.

Stephen Few - Perceptual Edge

Any interesting patterns in this set of numbers*?

** Example inspired by Cole's Storytelling with Data book/blog and Colin Ware's books.*

How many 7's are there?



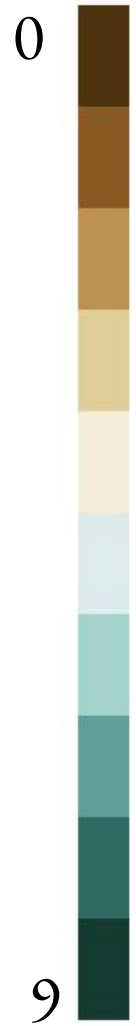
Let's try the typical default: add categorical colors

Let's try the typical default: add categorical colors

0->9 is ascending. Let's look at a sequential scale.

Viridis is also a popular colorscheme

Interested in extremes? Diverging from 4.5?

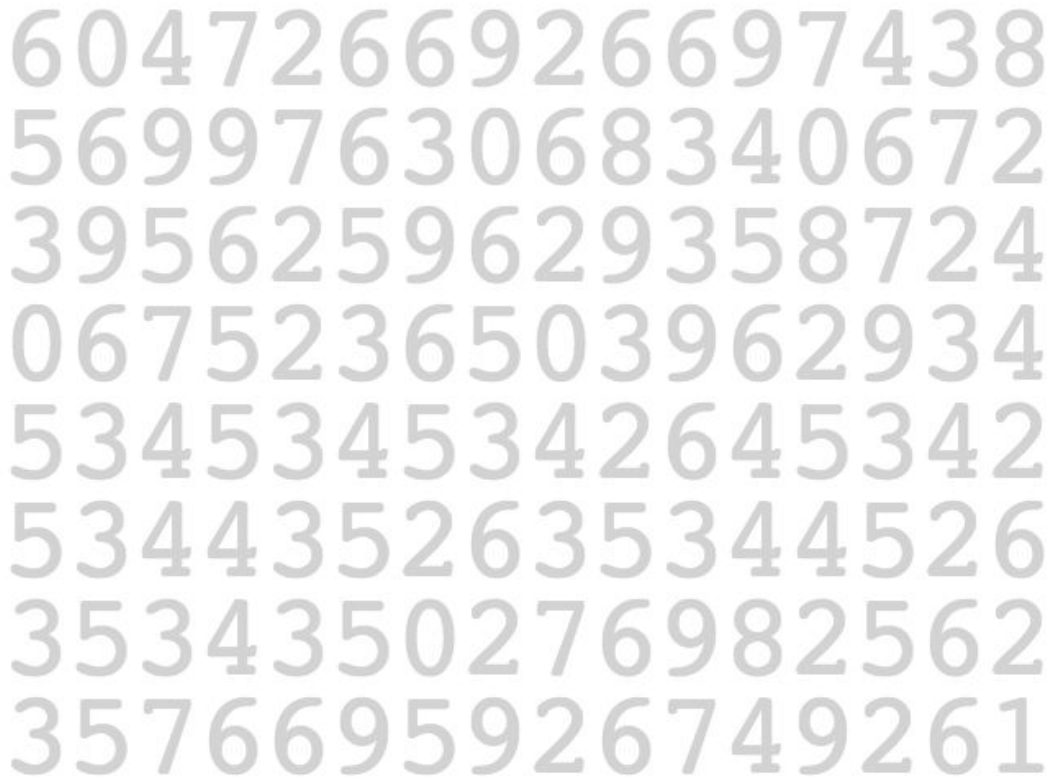


But, we know we want to count the number 7.

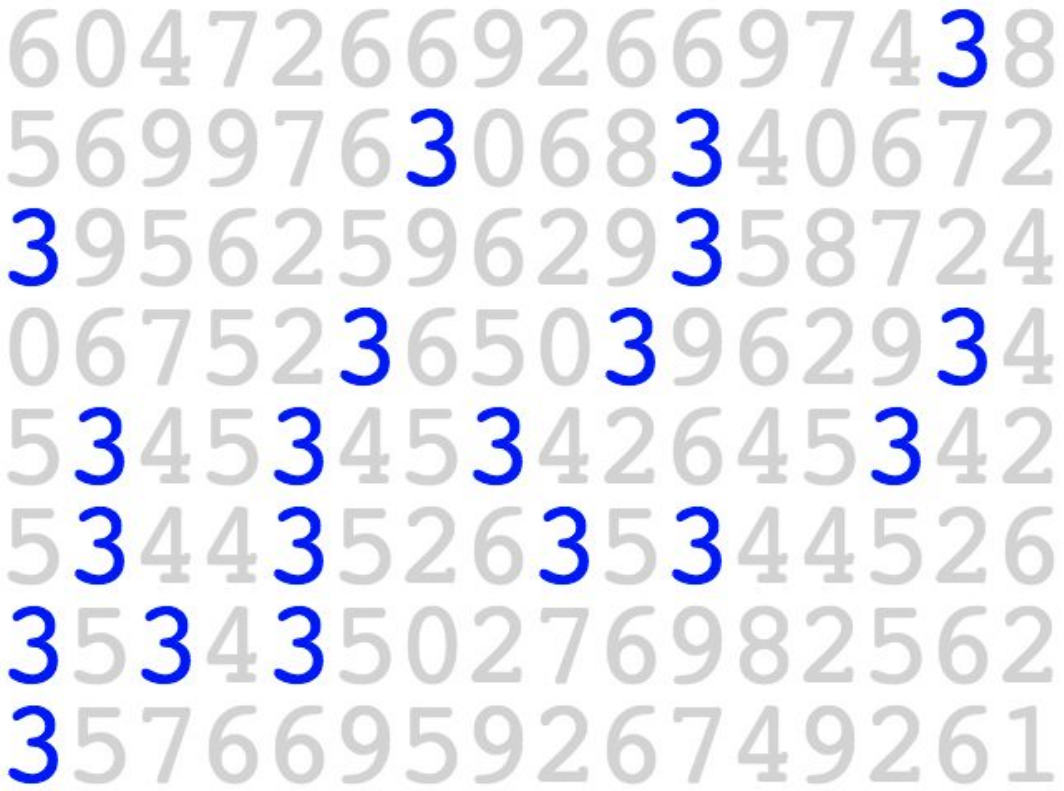
So, let's just show the 7s.

So, let's just show the 7s.

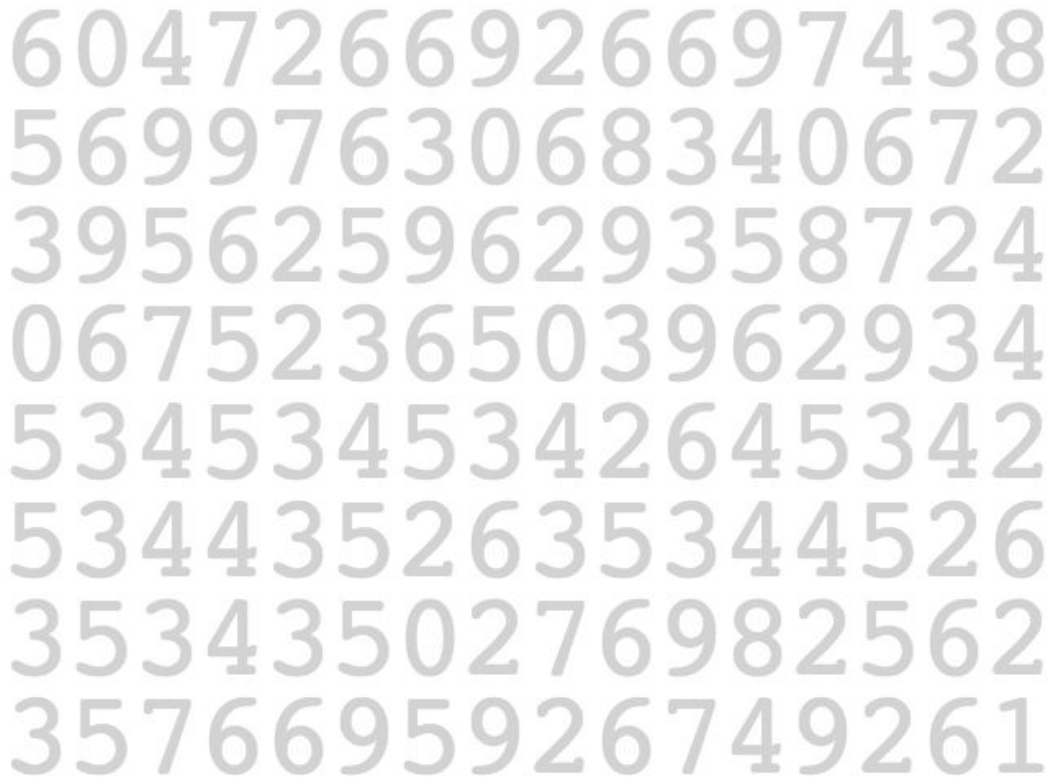
How many 3's are there?



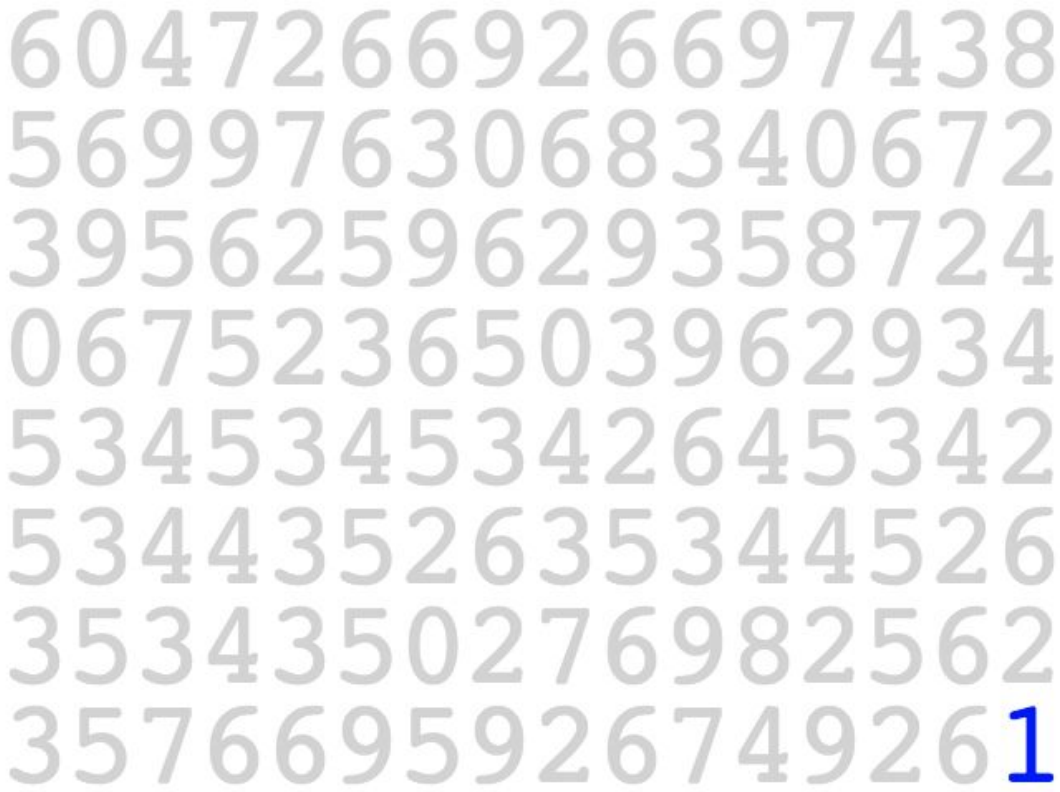
There are lots :)



What about 1's?



It is the loneliest number.



We can also use highlighting with "small multiples"

```
6047266926697438
5699763068340672
3956259629358724
0675236503962934
5345345342645342
5344352635344526
3534350276982562
3576695926749261
```

```
6047266926697438
5699763068340672
3956259629358724
0675236503962934
5345345342645342
5344352635344526
3534350276982562
3576695926749261
```

```
6047266926697438
5699763068340672
3956259629358724
0675236503962934
5345345342645342
5344352635344526
3534350276982562
3576695926749261
```

```
6047266926697438
5699763068340672
3956259629358724
0675236503962934
5345345342645342
5344352635344526
3534350276982562
3576695926749261
```

```
6047266926697438
5699763068340672
3956259629358724
0675236503962934
5345345342645342
5344352635344526
3534350276982562
3576695926749261
```

```
6047266926697438
5699763068340672
3956259629358724
0675236503962934
5345345342645342
5344352635344526
3534350276982562
3576695926749261
```

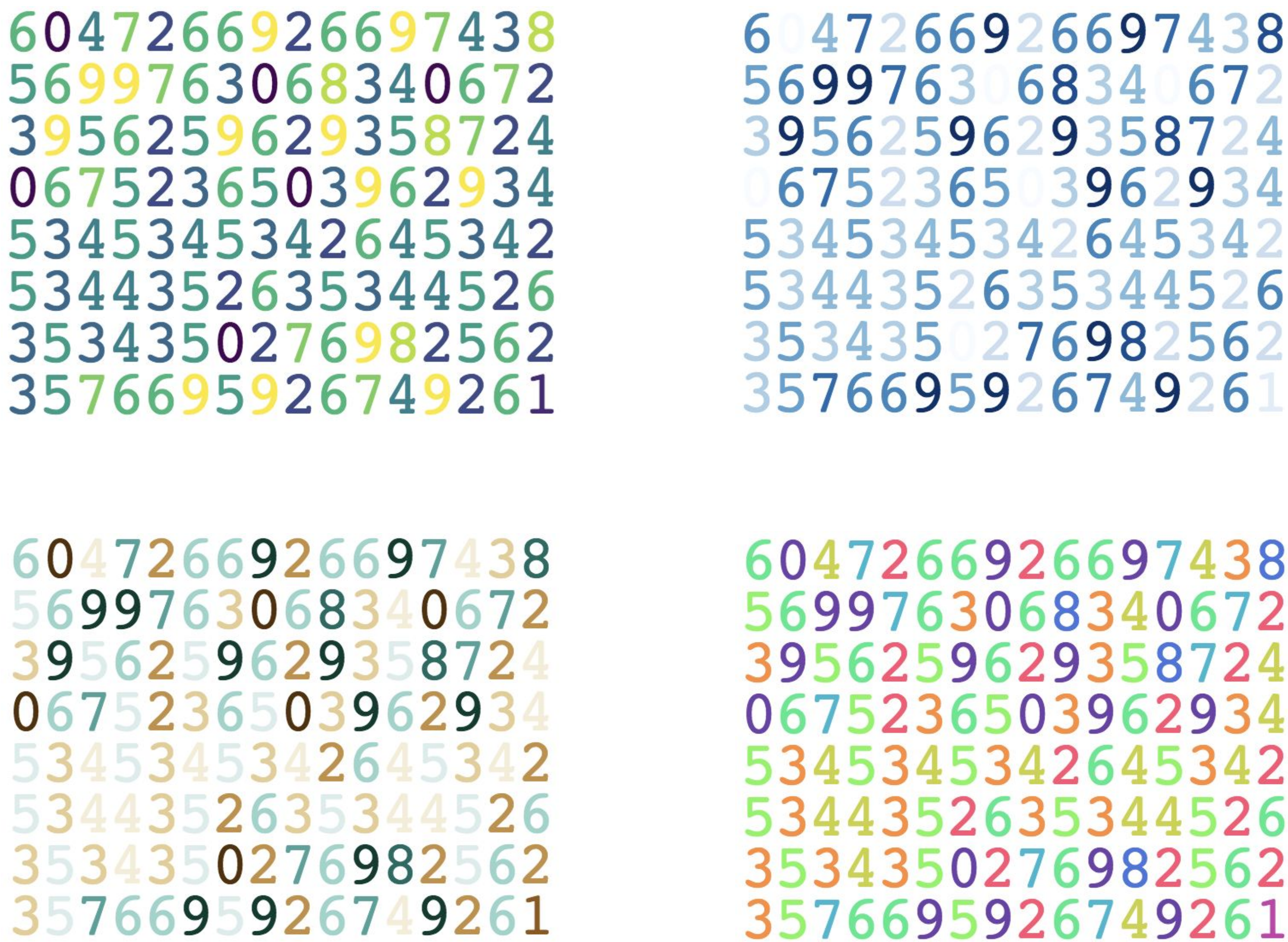
```
6047266926697438
5699763068340672
3956259629358724
0675236503962934
5345345342645342
5344352635344526
3534350276982562
3576695926749261
```

```
6047266926697438
5699763068340672
3956259629358724
0675236503962934
5345345342645342
5344352635344526
3534350276982562
3576695926749261
```

```
6047266926697438
5699763068340672
3956259629358724
0675236503962934
5345345342645342
5344352635344526
3534350276982562
3576695926749261
```

```
6047266926697438
5699763068340672
3956259629358724
0675236503962934
5345345342645342
5344352635344526
3534350276982562
3576695926749261
```

Or, for more general analysis, look at different color schemes that emphasize different aspects of the data.

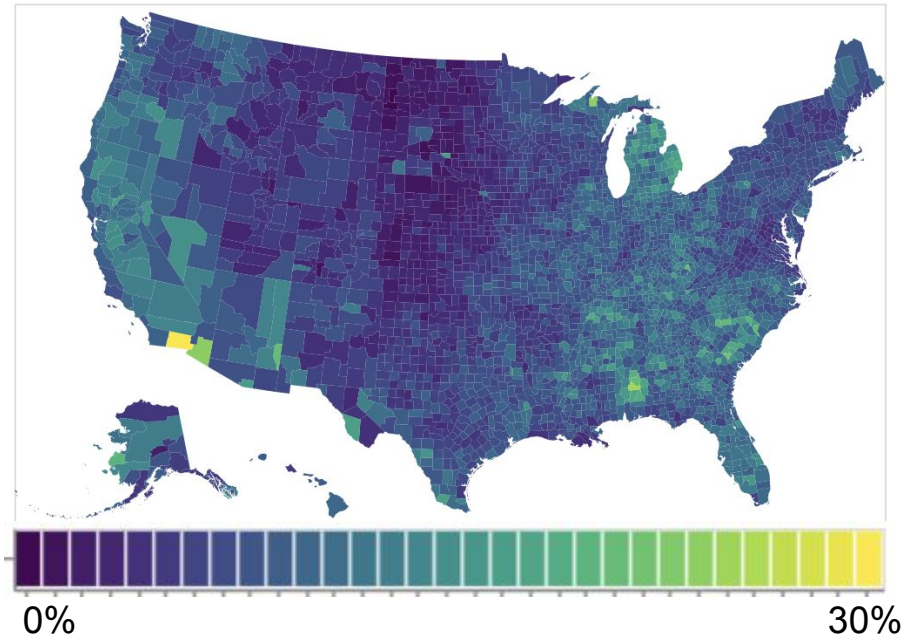


Use color intentionally

Make the mapping from numeric domain to color range be meaningful

Use color intentionally: Mapping from numbers to colors

Unemployment Data by County in the US, ~2009

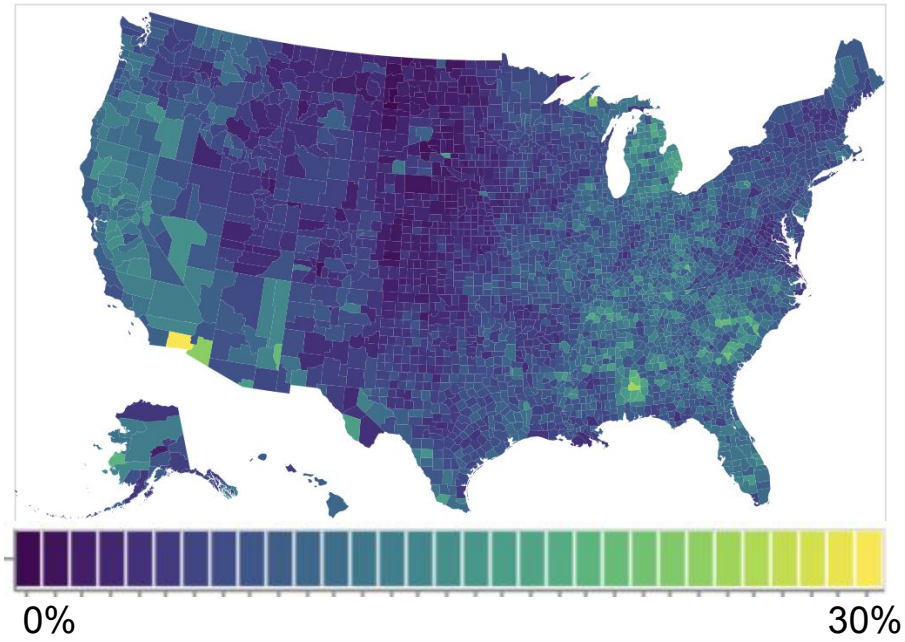


Altair Examples
Colab w/ this set of maps

Use color intentionally: Mapping from numbers to colors

Start with the default color mapping (Altair)

Typical Default Color Mapping
Color range spans from from 0 to max (30.1%).

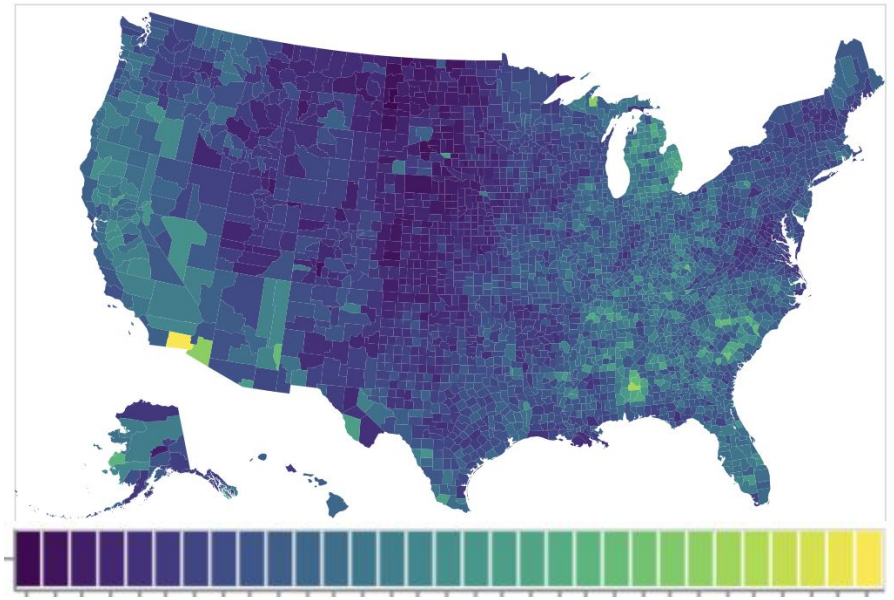


Altair Examples
Colab w/ this set of maps

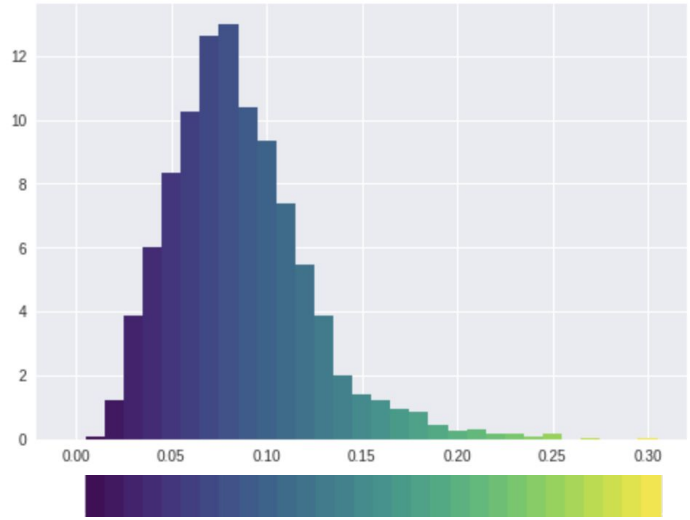
Use color intentionally: Mapping from numbers to colors

But, an extreme outlier overwhelms the map

Typical Default Color Mapping
Color range spans from from 0 to max (30.1%).



Distribution of unemployment rates by county in the dataset

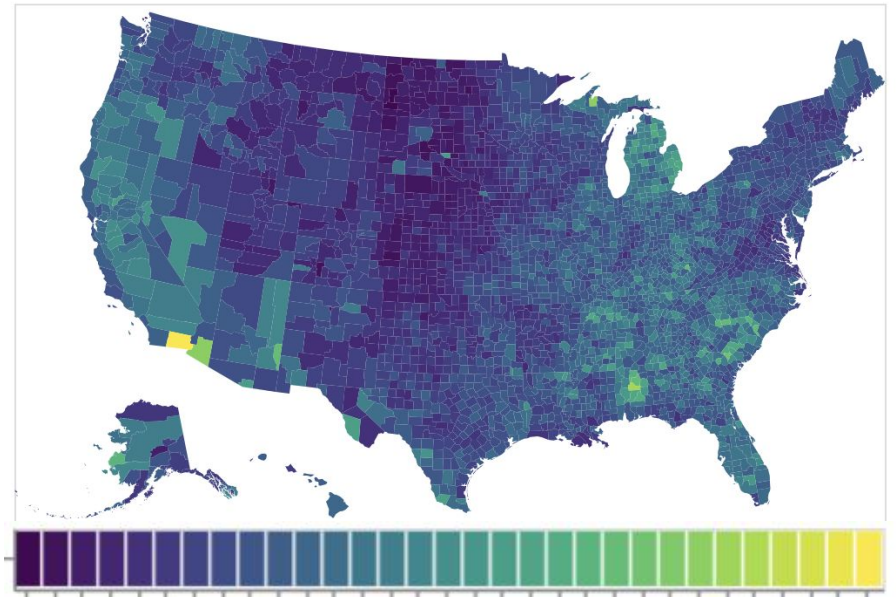


Altair Examples
Colab w/ this set of maps

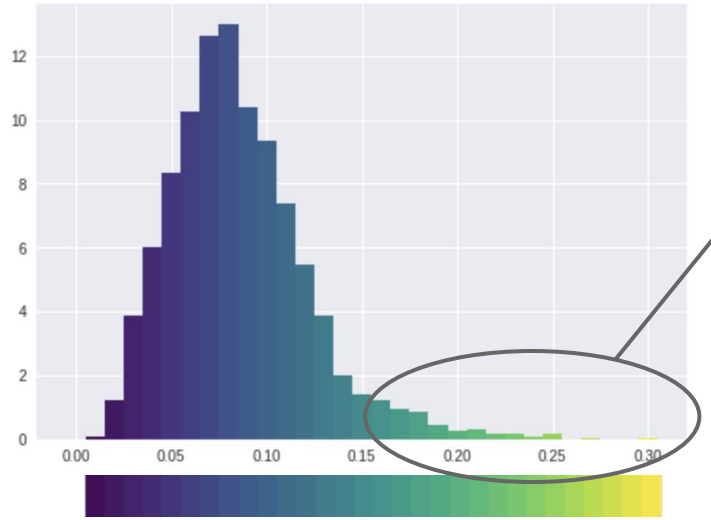
Use color intentionally: Mapping from numbers to colors

Harder to differentiate differences for bulk of the data

Typical Default Color Mapping
Color range spans from from 0 to max (30.1%).



Distribution of unemployment rates by county in the dataset



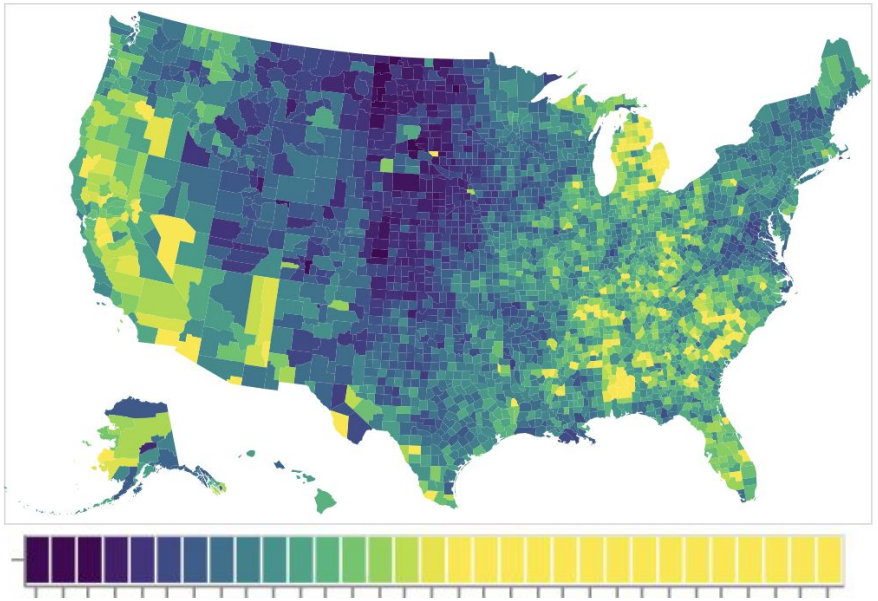
Only 4.5% of the counties have >15% unemployment, but we're using half our color scale on 15%-30%.

Altair Examples
Colab w/ this set of maps

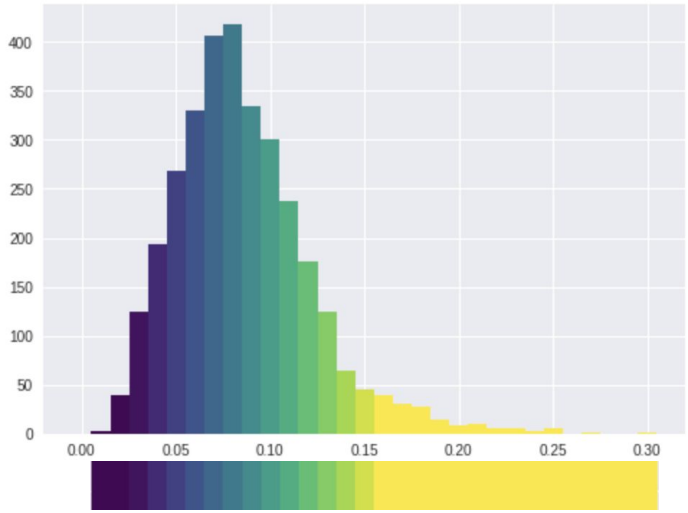
Use color intentionally: Mapping from numbers to colors

Use distribution of the data, rather than max value, to set color min/max

Intentional Range, based on the data distribution
Color range is set to min of 2% and max of 15.5%.



Distribution of unemployment rates by county in the dataset



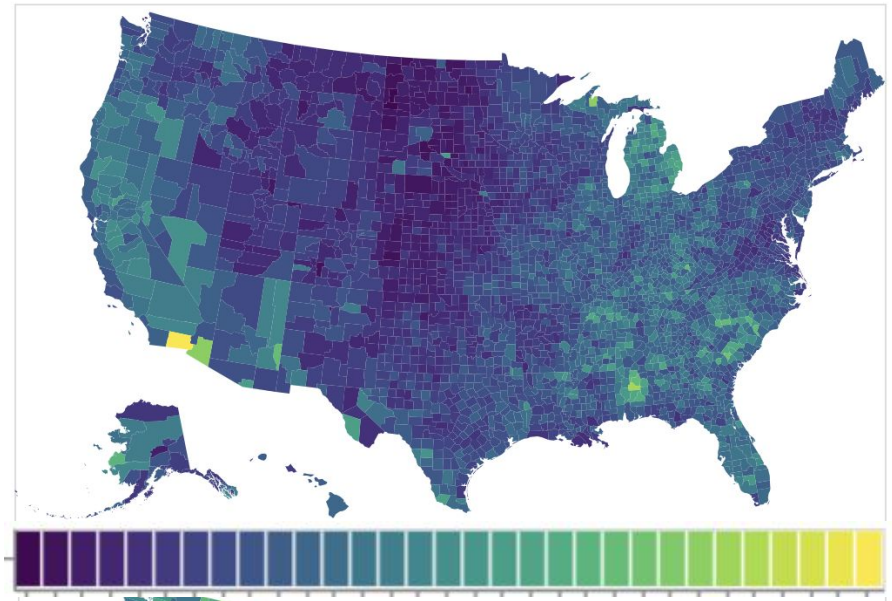
More than doubling the perceptual differentiation, with little loss on the extremes.

Altair Examples
Colab w/ this set of maps

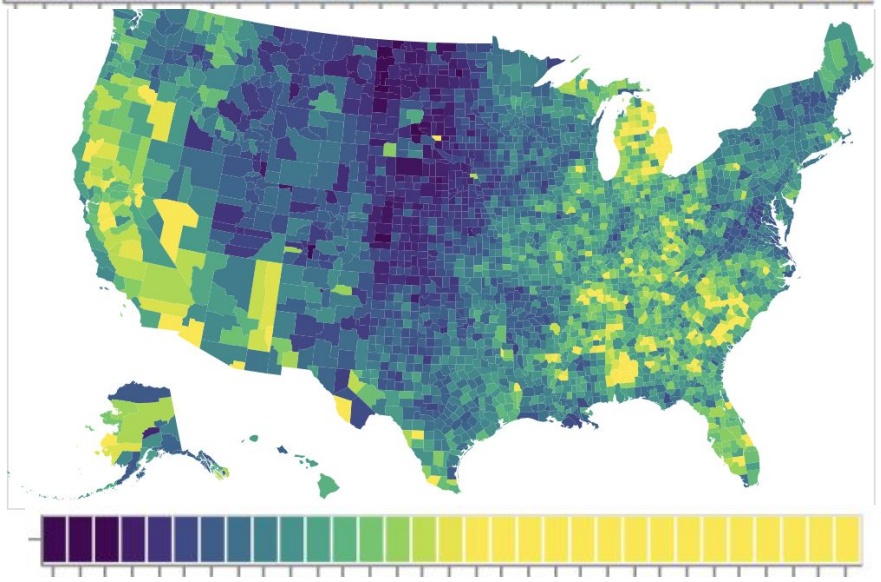
Use color intentionally: Mapping from numbers to colors

Let's compare.

Typical Default Color Mapping
Color range spans from from min to max (27%).



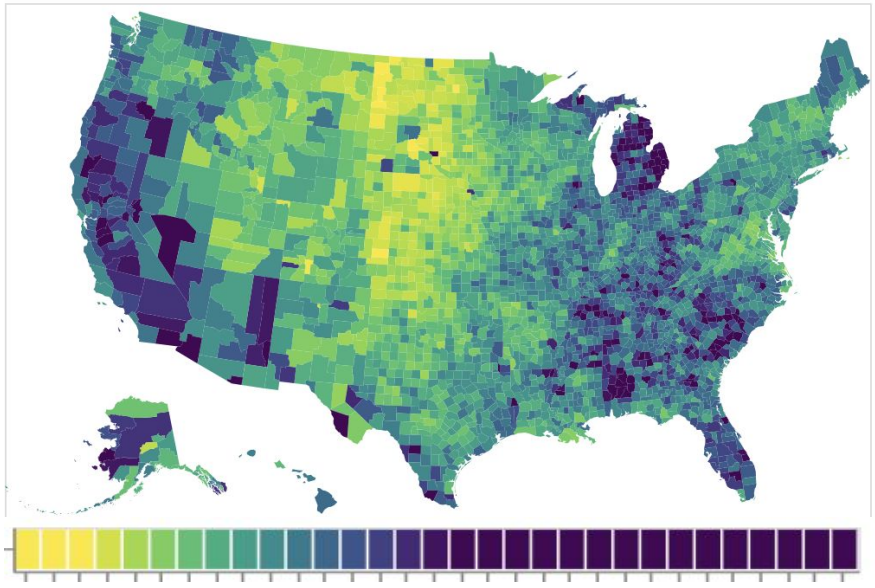
Intentional Range, based on data distribution
Color range is set to min of 2% and max of 15.5%.



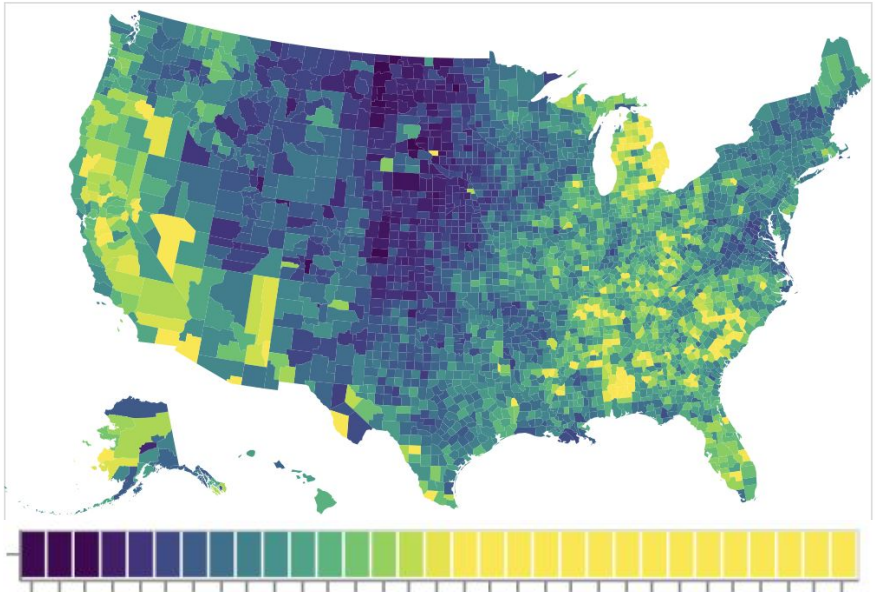
Use color intentionally: Mapping from numbers to colors

What if we flip the color scheme?

Intentional Range, flip the color map
Lowest unemployment rates are yellow, highest rates are purple.



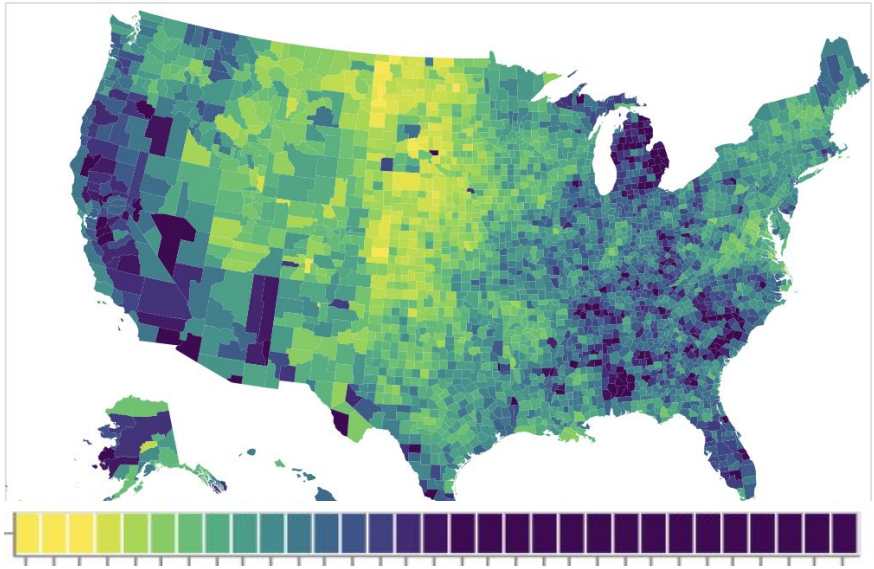
Intentional Range
Highest unemployment rates are yellow, lowest are purple.



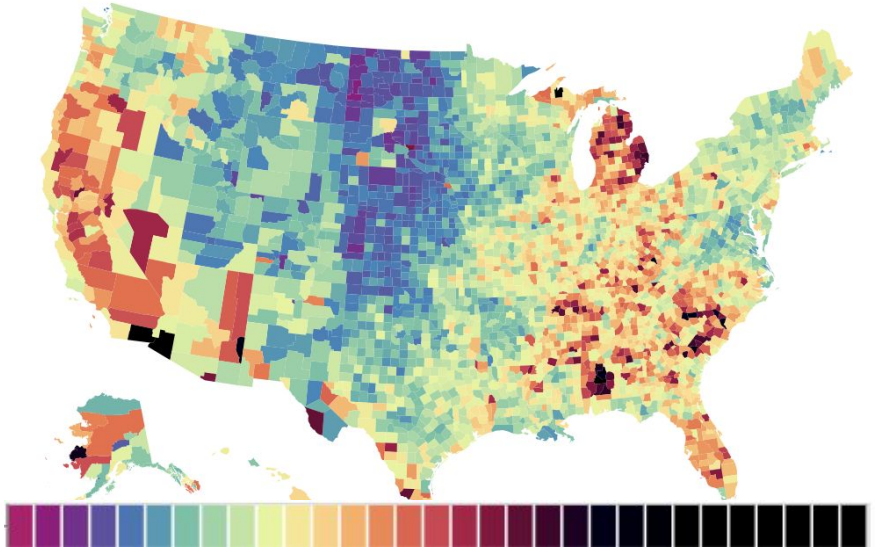
Use color intentionally: Mapping from numbers to colors

Or try a different color scheme?

Intentional Range, flip the color map
Lowest unemployment rates are yellow, highest rates are purple.



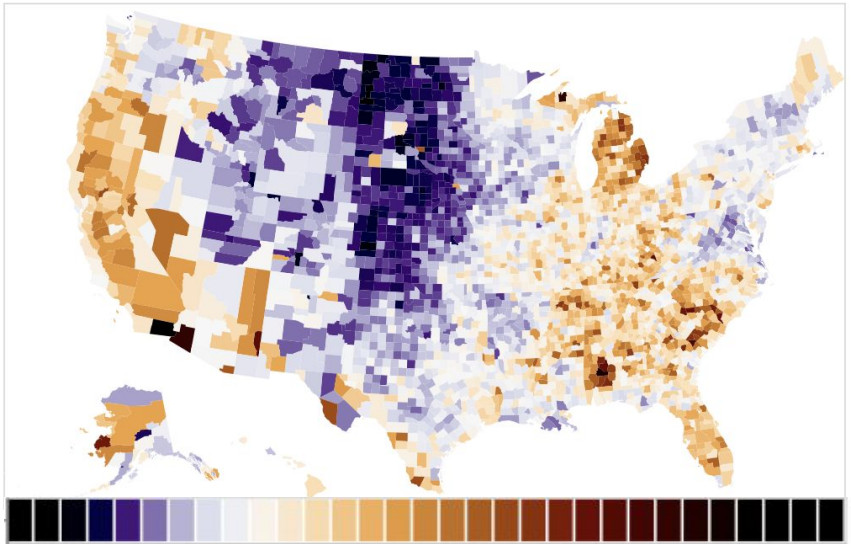
Another colorscheme?
Do we notice the same things? Different things?



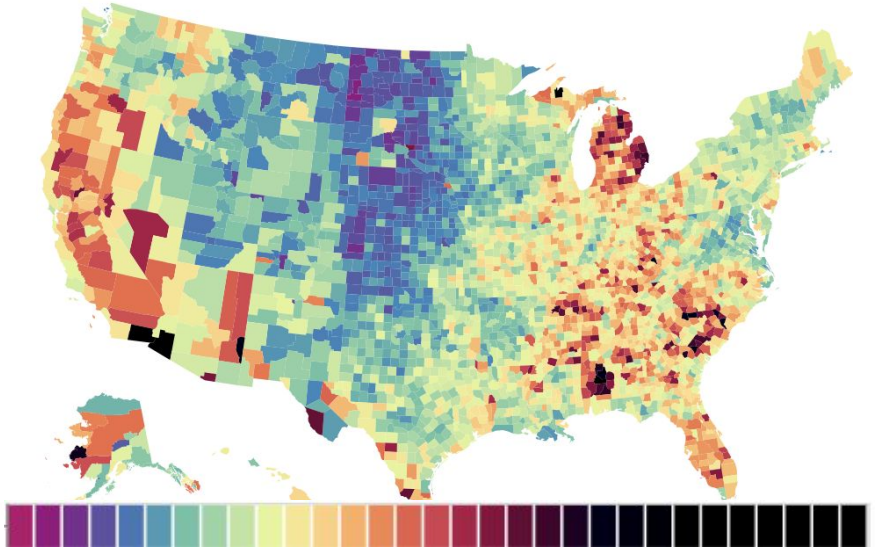
Use color intentionally: Mapping from numbers to colors

Perhaps diverging is more appropriate? Compare to the median.

Diverging
What if we intentionally compared to the median value? Purple is below 8.5%, median is shown as white, above the median as orange.



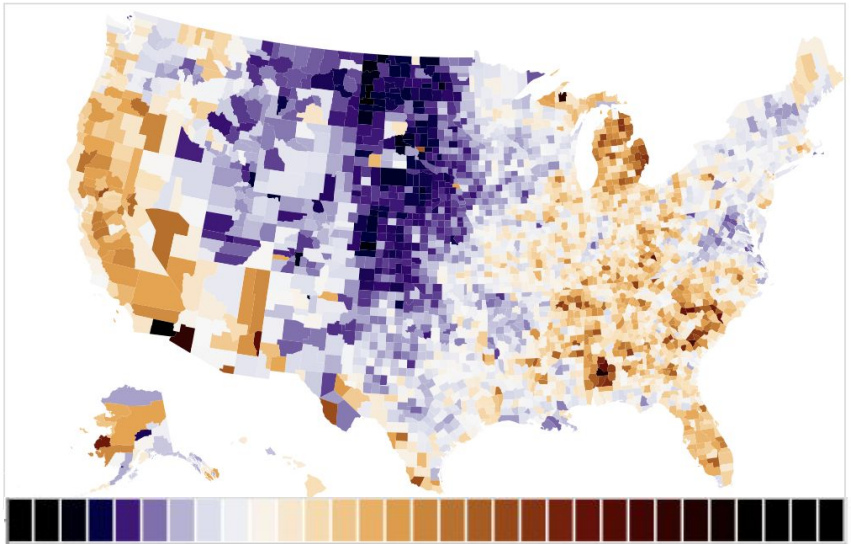
Another colorscheme?
Do we notice the same things? Different things?



Use color intentionally: Mapping from numbers to colors

But, is the 8.5% median unemployment rate the most relevant "middle"? What about the "ideal" rate of 5.5% for a healthy economy?

Diverging
What if we intentionally compared to the median value (8.5%)?



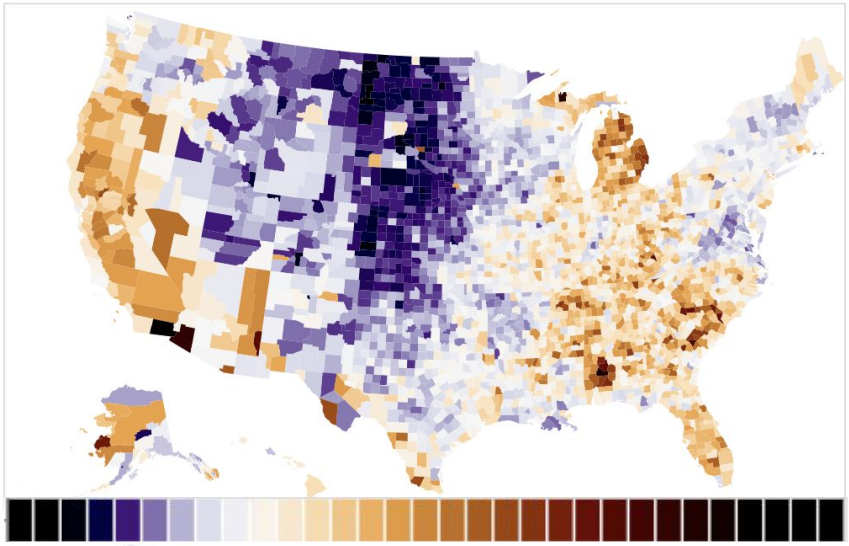
Very dark extreme ends start to look similar.

We could "clamp" to a mid-dark max value to prevent this.

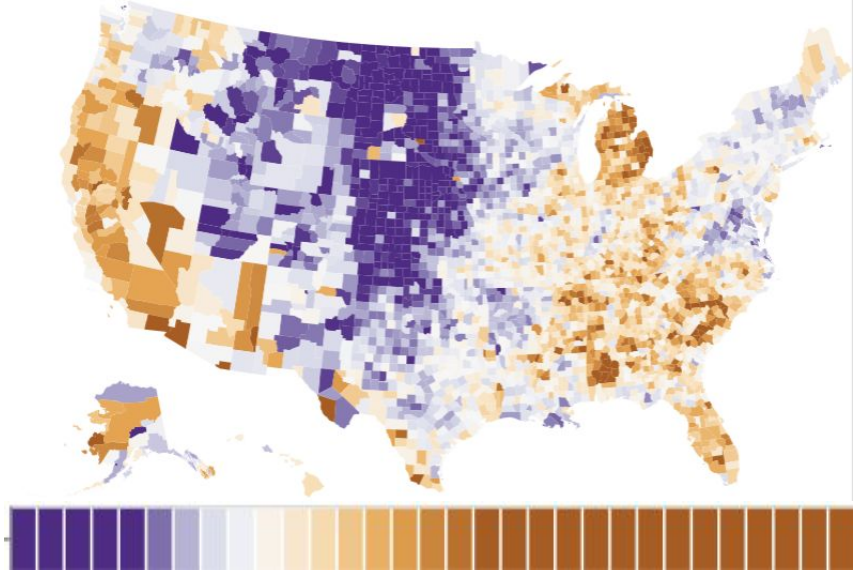
Use color intentionally: Mapping from numbers to colors

But, is the 8.5% median unemployment rate the most relevant "middle"? What about the "ideal" rate of 5.5% for a healthy economy?

Diverging
What if we intentionally compared to the median value (8.5%)?



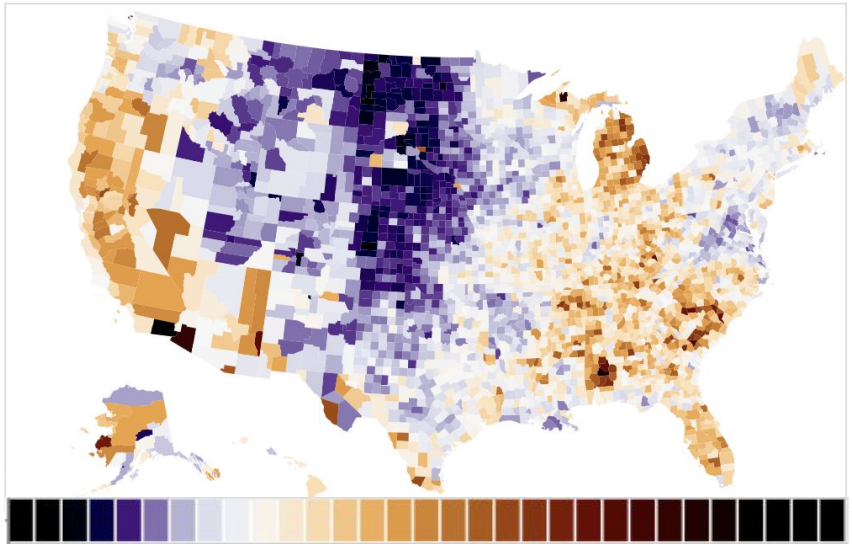
Diverging
From median, Clamped min/max values.



Use color intentionally: Mapping from numbers to colors

What should we set as the middle value?

Diverging
What if we intentionally compared to the median value (8.5%)?

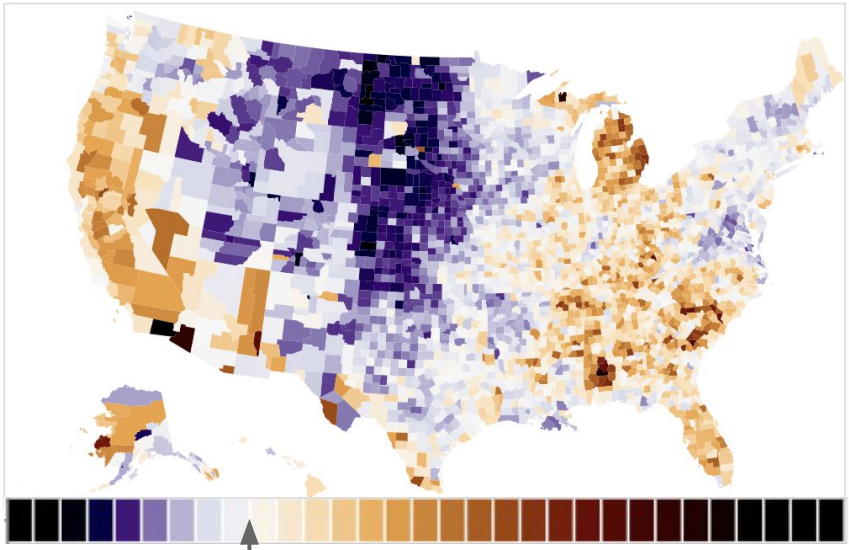


Why should the "middle" be 8.5%?

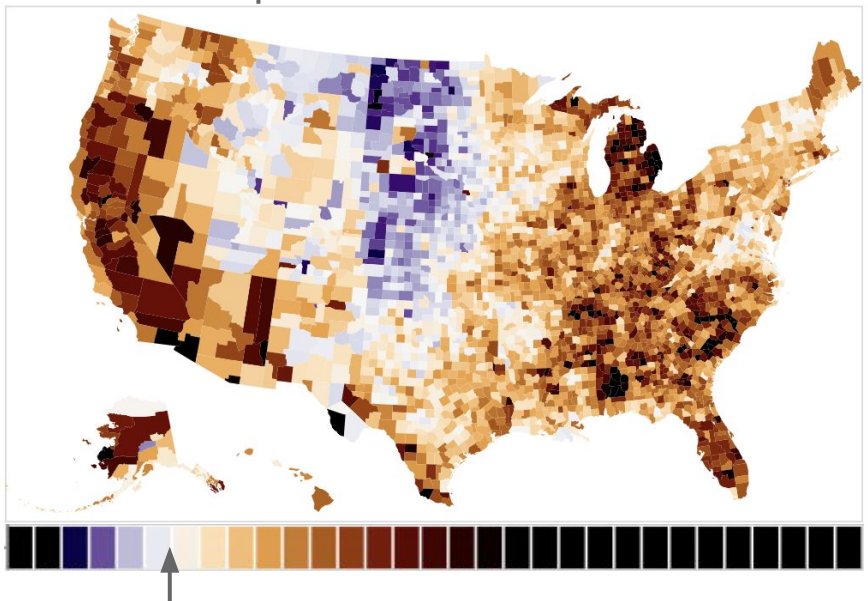
Use color intentionally: Mapping from numbers to colors

But, is the 8.5% median unemployment rate the most relevant "middle"? What about the "ideal" rate of 5.5% for a healthy economy?

Diverging
What if we intentionally compared to the median value (8.5%)?



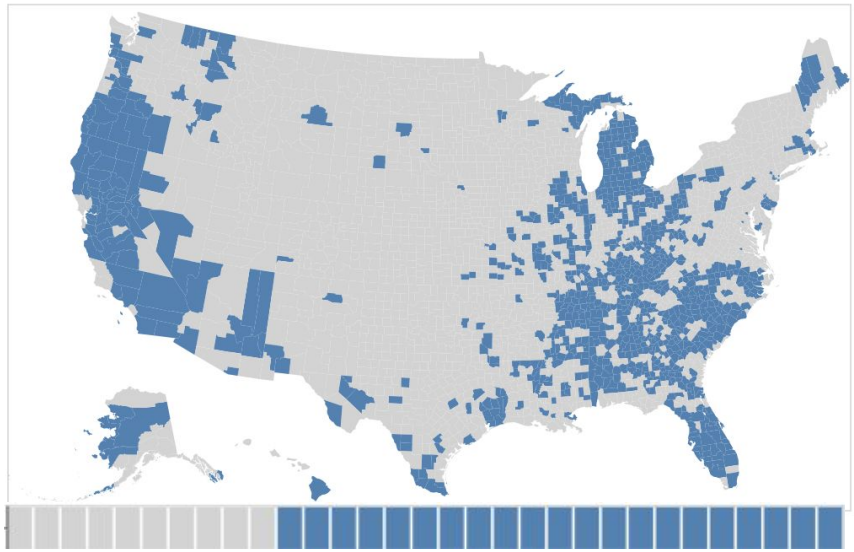
Diverging:
compare to ideal value according to economists (~5.5%) rather than median



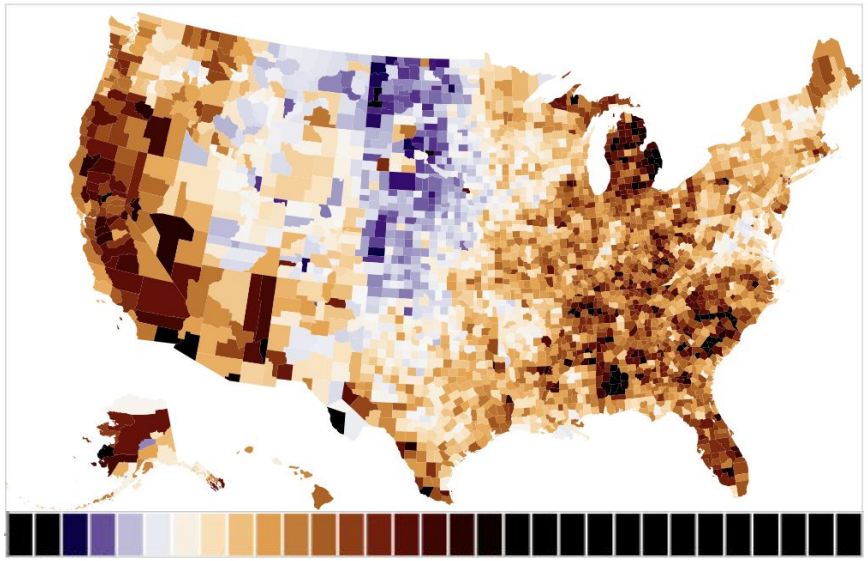
Use color intentionally: Mapping from numbers to colors

We could put in a hard cut-off, say to highlight areas with greater than 10% unemployment.

Highlight
Just focus on areas with >10% unemployment (in blue).



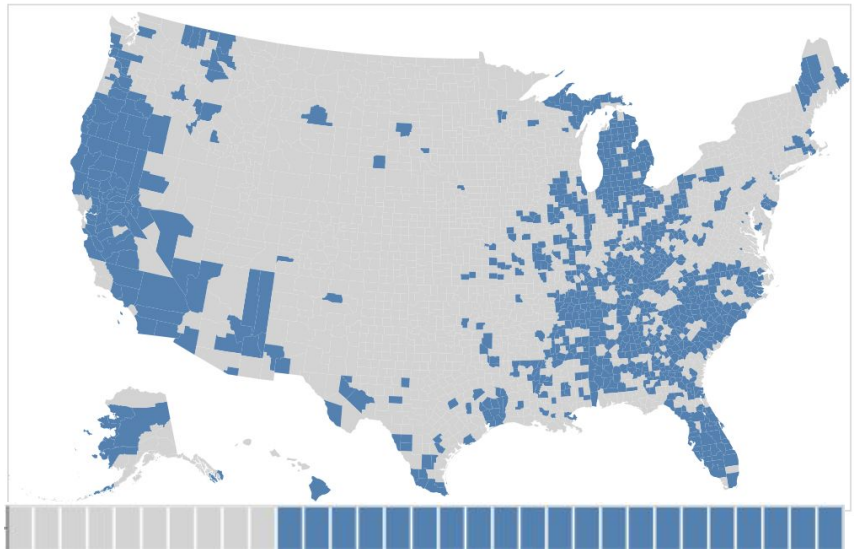
Diverging:
compare to ideal value according to economists (~5.5%) rather than median



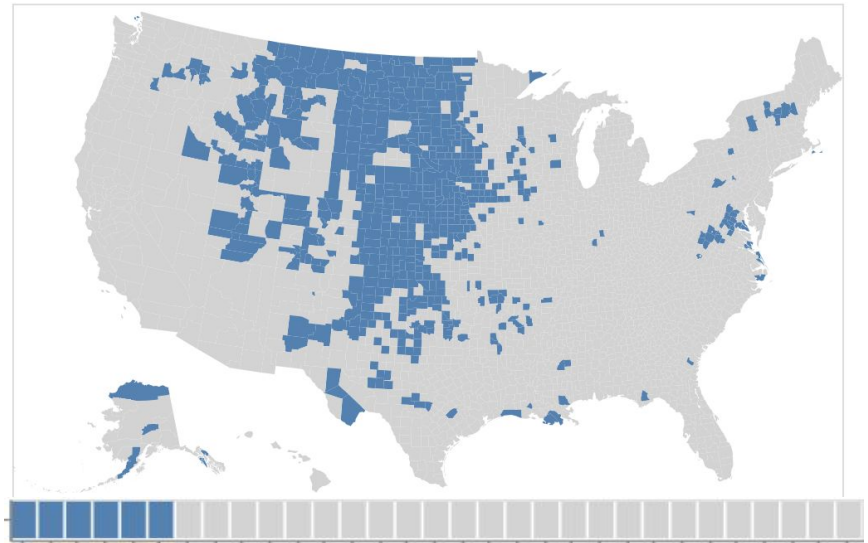
Use color intentionally: Mapping from numbers to colors

Or, a hard cut-off of <6%.

Highlight
Just focus on areas with >10% unemployment (in blue).

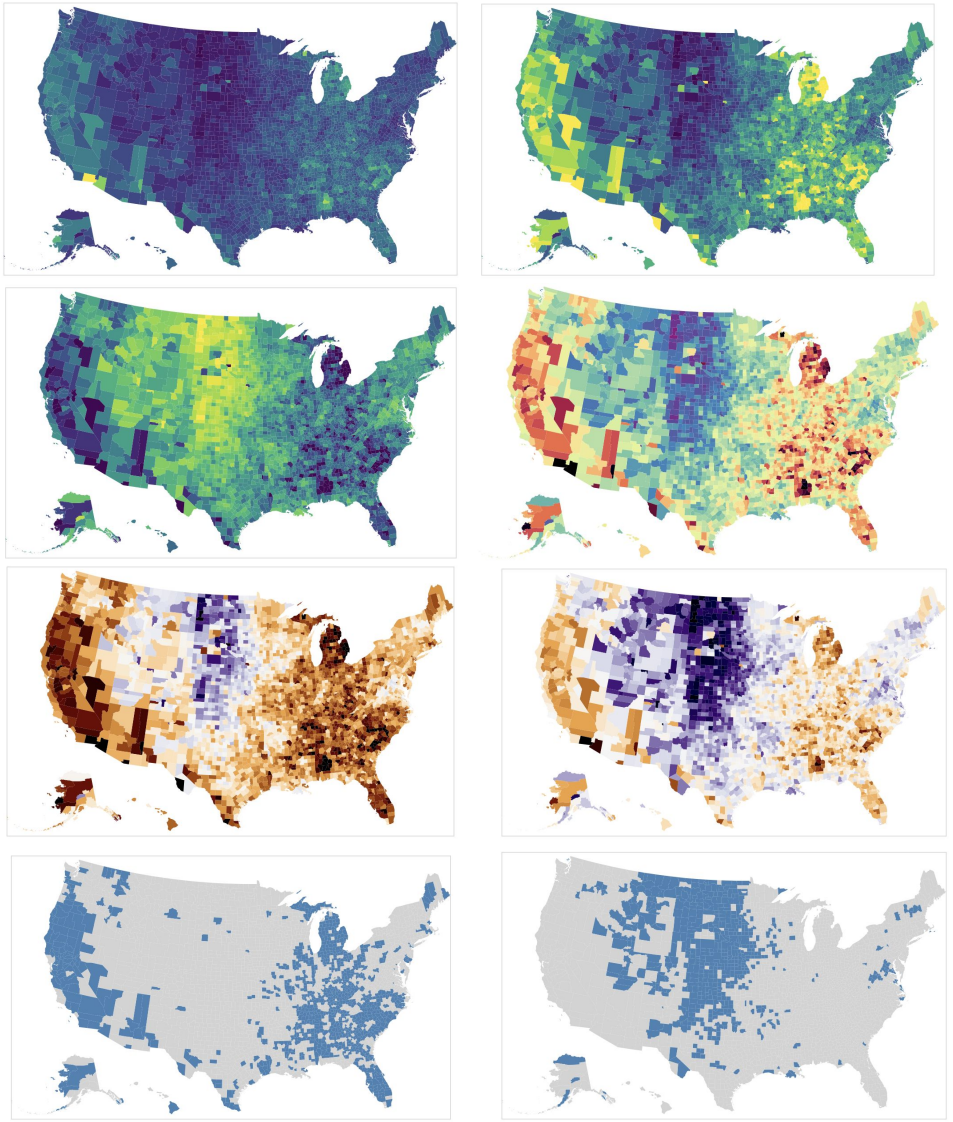


Highlight:
Focus on areas with <6% unemployment



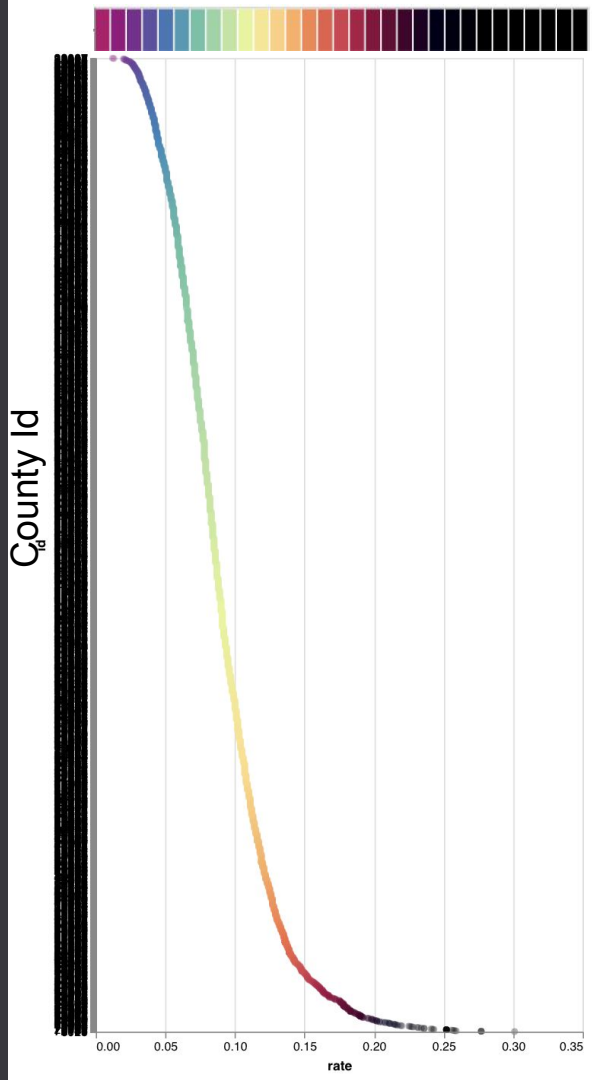
Use color intentionally: Mapping from numbers to colors

Different mappings ask different questions of our data. The "right" color mapping depends on what question you want to ask your data (like parameters/algorithms in stats/ML).

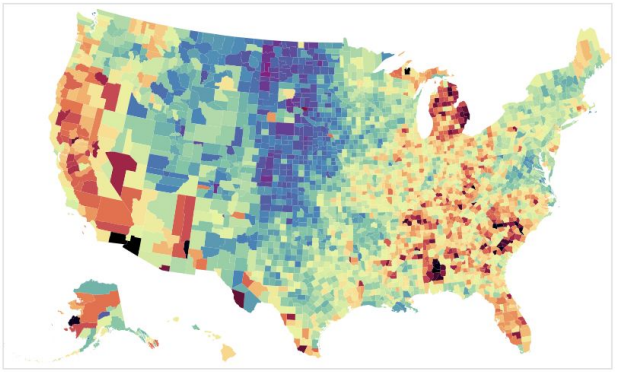


Using colormaps intentionally:

If you want to optimize for **quantitative** rather than qualitative comparisons; try a chart that relies on position rather than color.

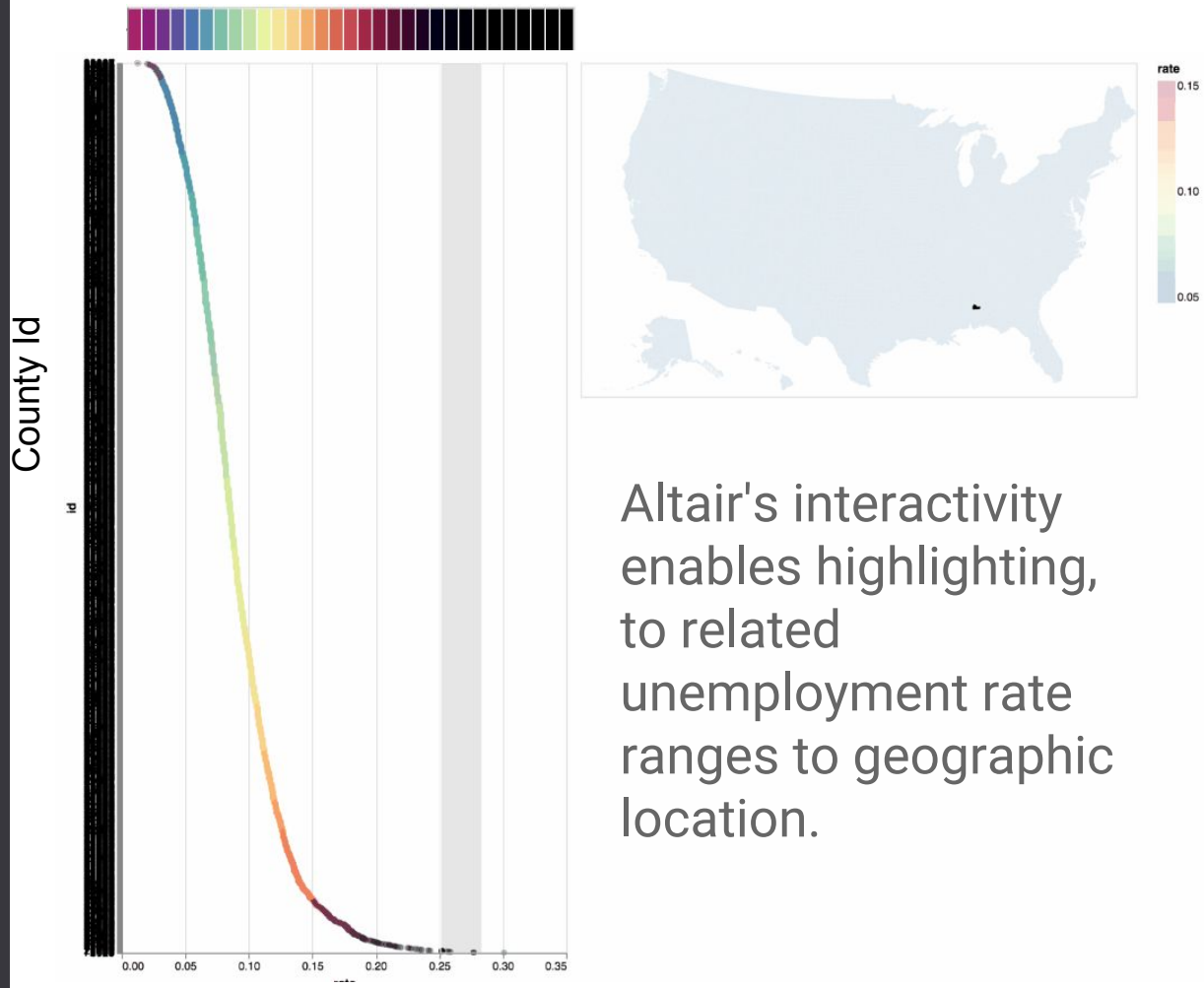


Unemployment Rate



Using colormaps intentionally:

If you want to optimize for **quantitative** rather than qualitative comparisons; try a chart that relies on position rather than color.

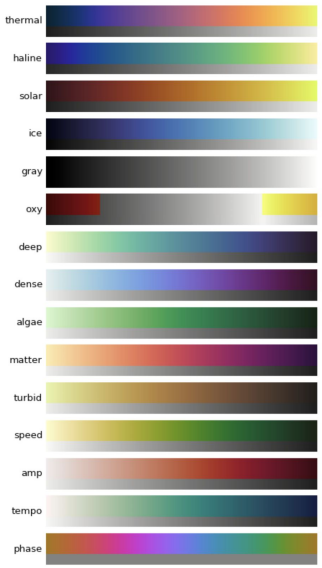


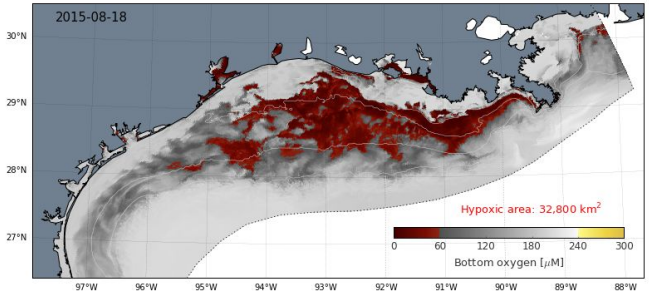
Altair's interactivity enables highlighting, to related unemployment rate ranges to geographic location.

Use color intentionally

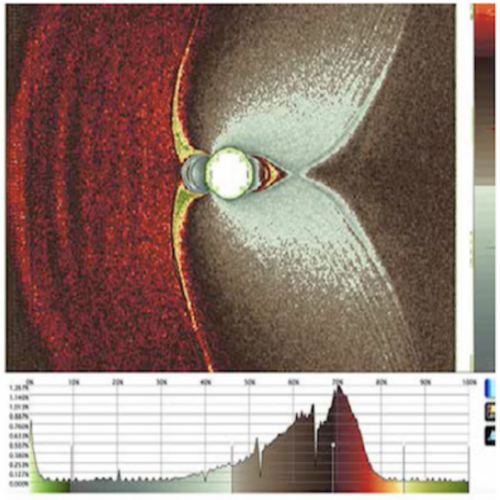
Resources

cmocean color scales





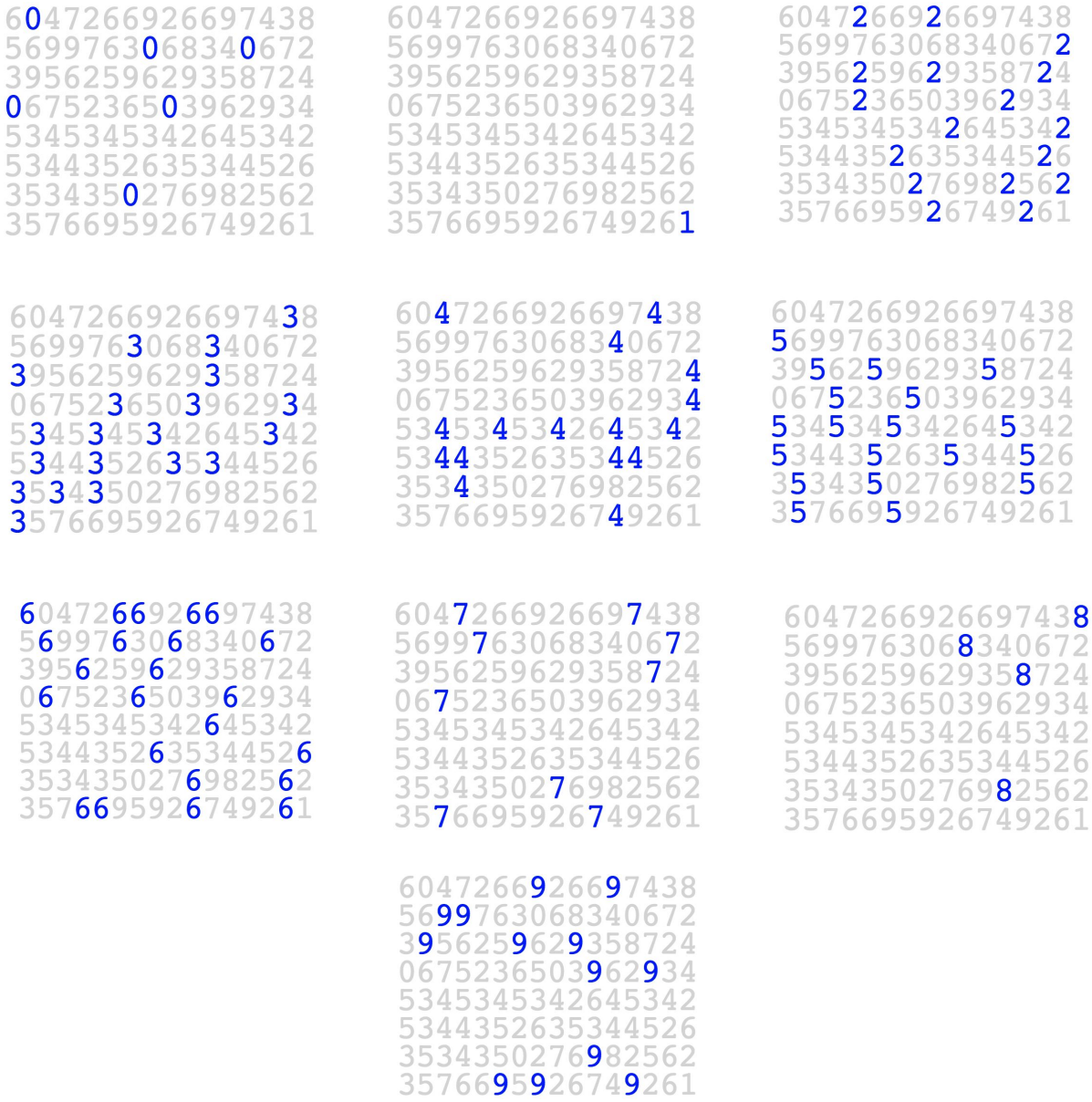
sciviscolor



Use color intentionally

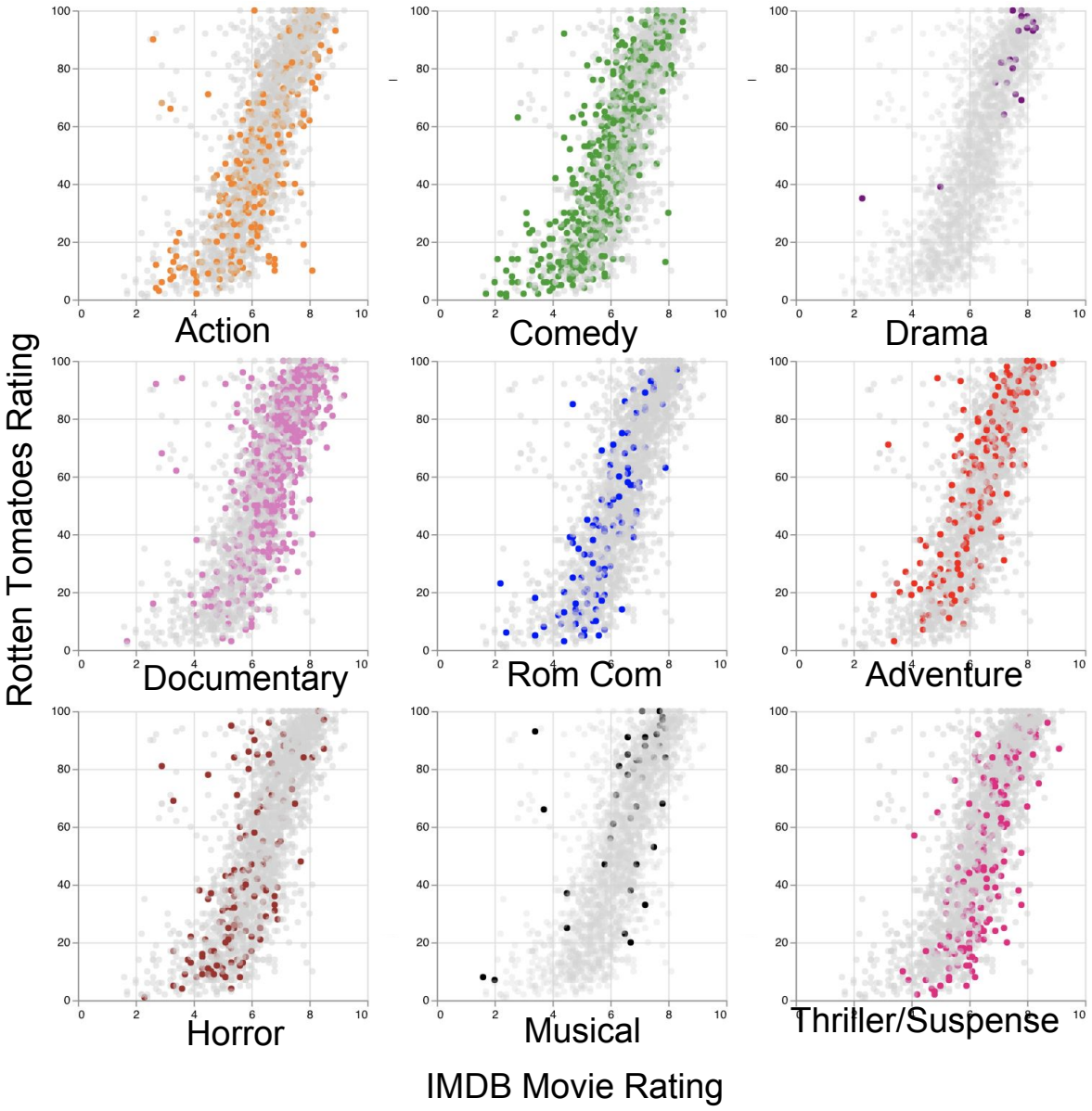
With a highlight color to focus attention

Use color intentionally



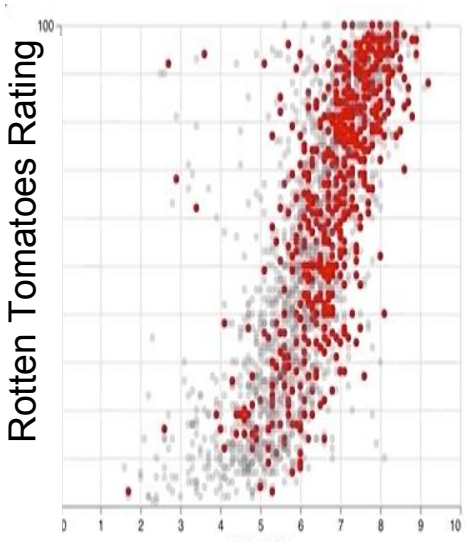
Use color intentionally

Interactivity example

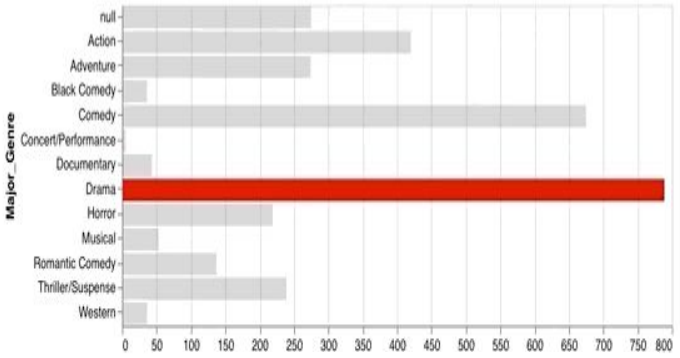


Use color intentionally

Interactivity example



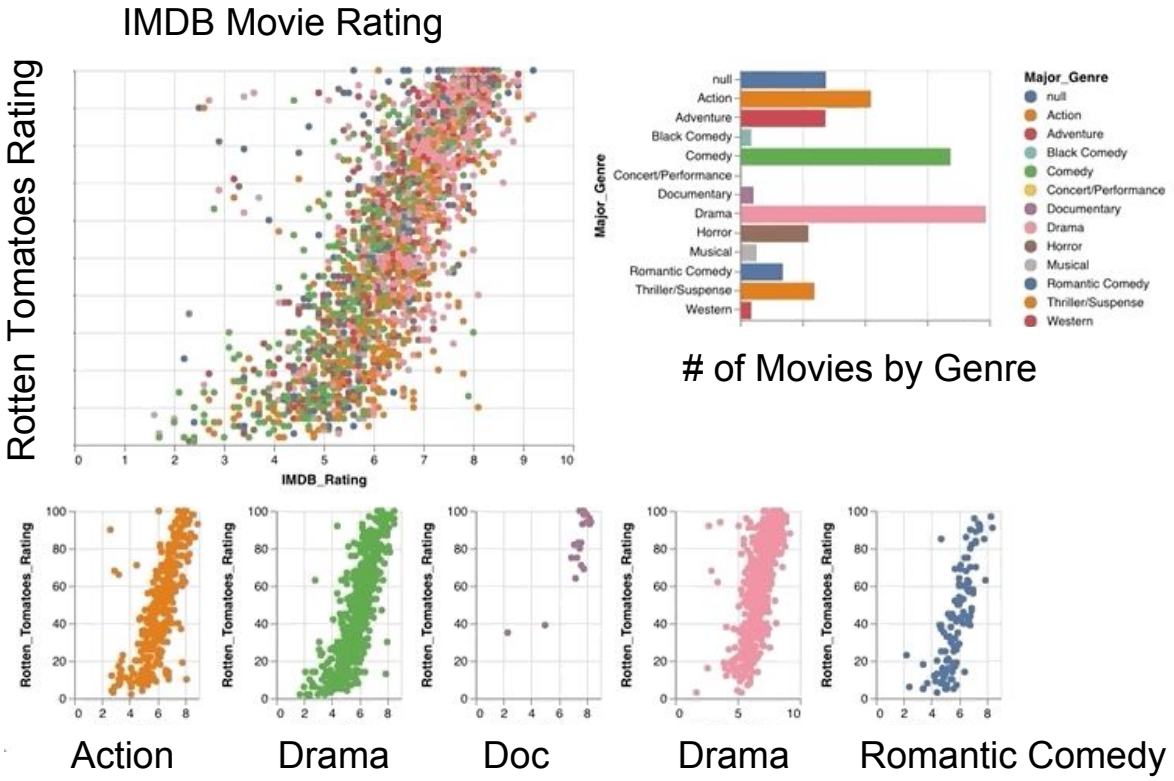
IMDB Movie Rating



# of Movies by Genre

Use color intentionally

Interactivity example



Optimize for your time, energy, and attention

Optimize for time, energy, attention

In science, we often have "data processing pipelines" (processes you run regularly on similar types of data).

Optimize for time, energy, attention

In science, we often have "data processing pipelines" (processes you run regularly on similar types of data).

Want to know:

Is the data ok? Broken?

Is there anything "interesting" here?

Optimize for time, energy, attention

In science, we often have "data processing pipelines" (processes you run regularly on similar types of data).

Want to know:

- Is the data ok? Broken?
- Is there anything "interesting" here?

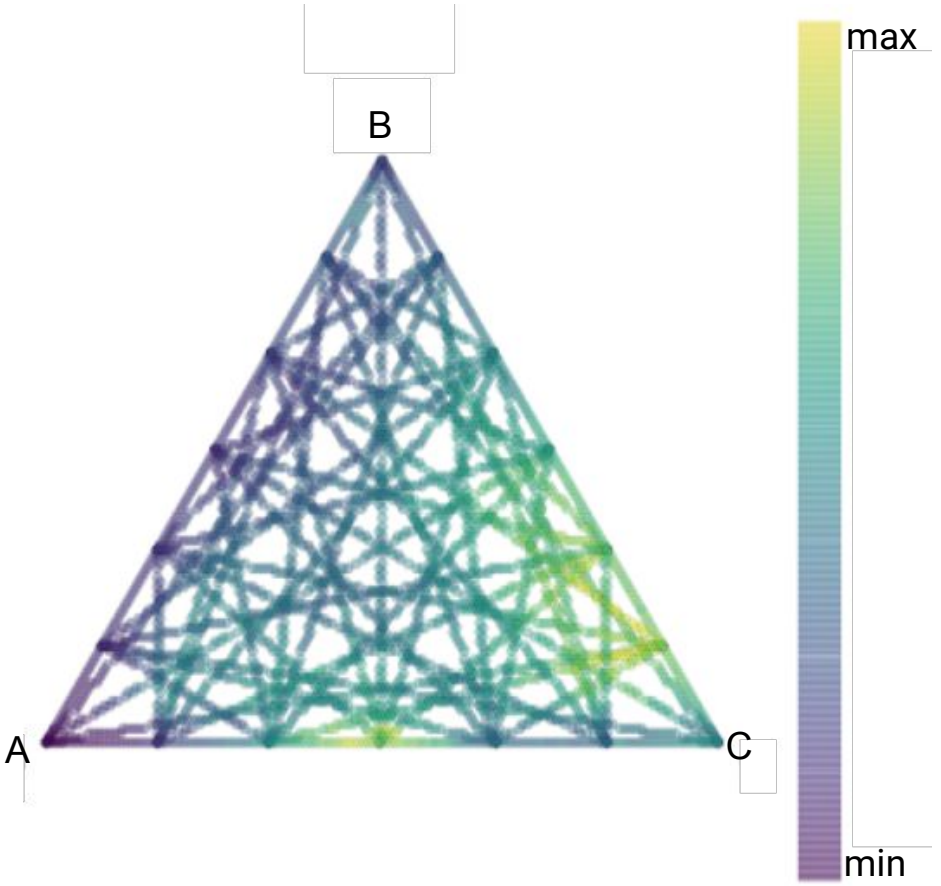
Create:

- Many charts
- Mostly boring
- Important stuff is obvious

Optimize for time, energy, attention

Case Study - Collaboration w/ Google Accelerated Science & CalTech

Looking for areas of interest (local min, max, slope changes) in 3-part mix

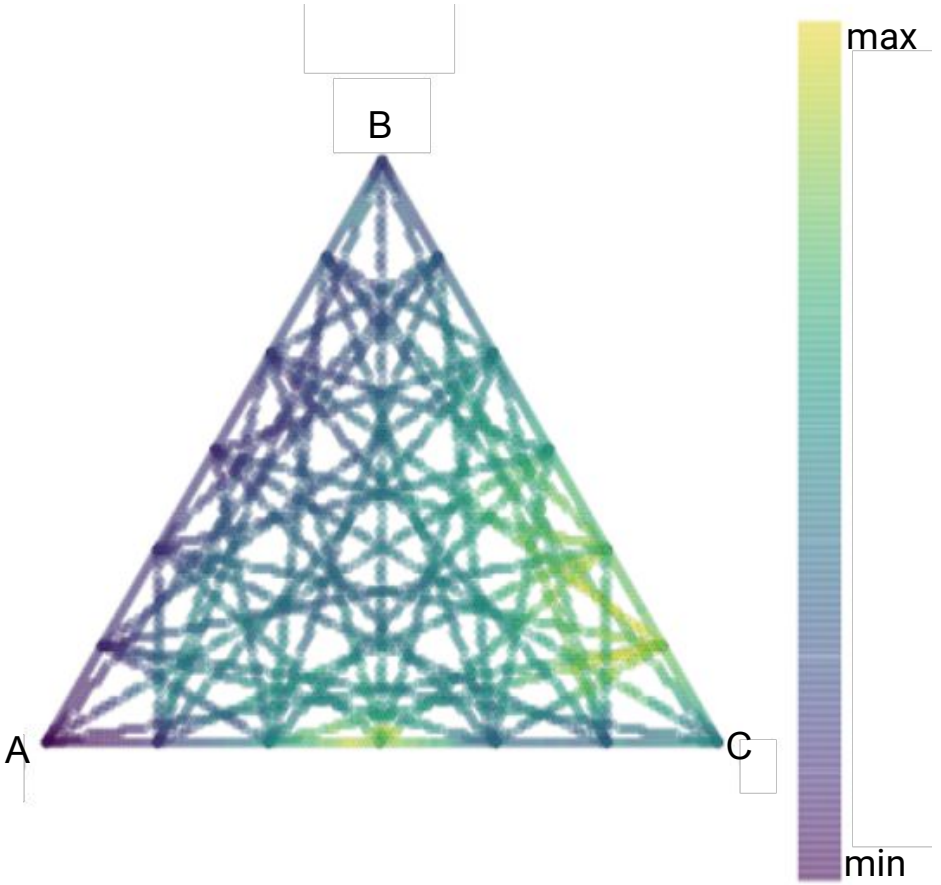


Optimize for time, energy, attention

Issues w/ original:

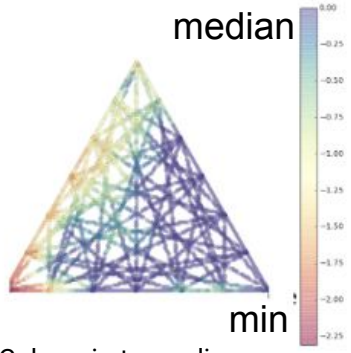
1 - Max values more eye-catching than min, middle. Hard to see "interesting" in mid-colorscale.

2 - Extreme values wash out the color scale

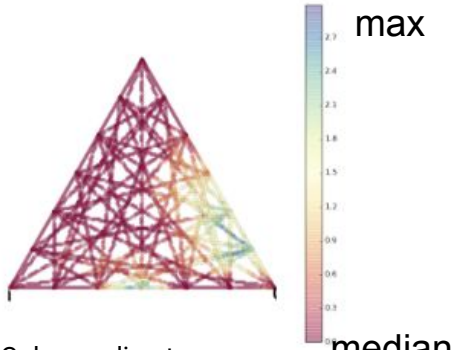


Optimize for time, energy, attention

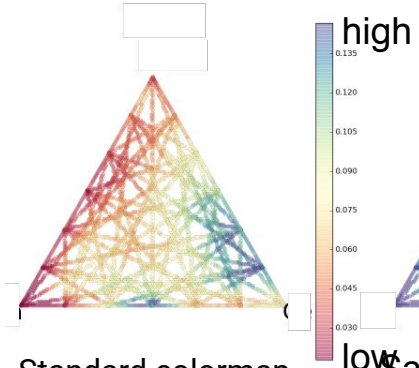
Small mulitiples; Intentional color Optimize for seeing if something is important, quickly



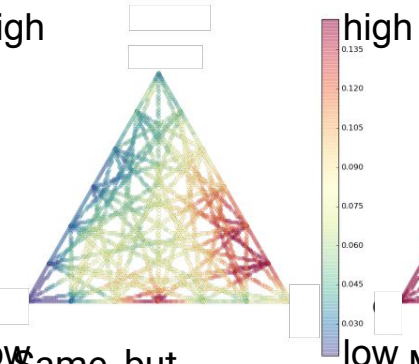
Color min-to-median, Everything higher is purple



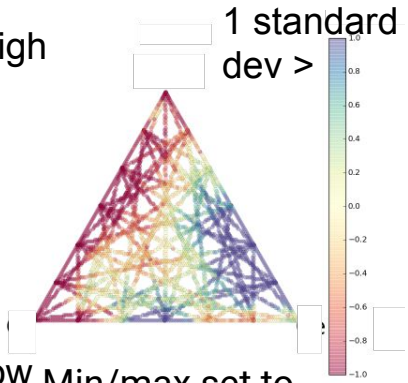
Color median-to-max, everything lower is red.



Standard colormap with lots of hue variation,
max to min set by min/max in dataset



Same, but inverted min/max colors



Min/max set to first standard deviation

In summary:

Is it good or bad depends on your goals, context, and constraints.

In summary:

Create small multiples (many, related charts)

Use color intentionally

Optimize your charts for your time/energy/attention

Resources

- cmocean color scales
- sciviscolor
- Varying binwidths and histograms
- Colin Ware's Information Visualization book
- Cole Nussbaumer - Storytelling with Data

Examples from the talk

- colab with geographic maps/colors
- Colab with small multiples