Introduction to Data Visualization

Govern for America

Jake Rozran

June 24, 2022

Who Am I?

- Please call me Jake (I'll accept Jake from State Farm)
- I am an Adjunct Professor of Data Science and Statistics @ Villanova University
- Data Science Practice Lead at a company called CivicActions
- Dad; Data Nerd; Philadelphian
- Please connect with me! www.jakelearnsdatascience.com

Where is this "Expertise" Coming From?

Where Else is this "Expertise" Coming From?

It's Friday afternoon and y'all have been at this for a few days now... I am going to try to make this as interactive as possible. Please stop me to ask questions at any point!

So... Let's Get Setup for the First Exercise

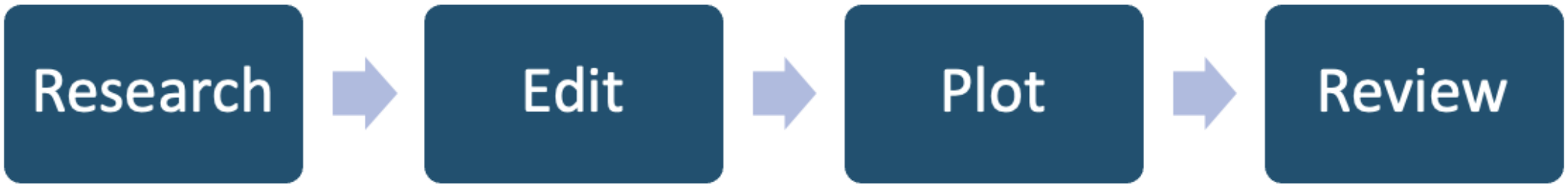
Download and open, in Excel, the Iris dataset OR

- Only if you already have Microsoft Excel installed!
- https://tinyurl.com/GFA-data-download

Open the data in Google Sheets

- Make a copy of the spreadsheet so you can edit!
- https://tinyurl.com/GFA-data-google-sheets

How to Create Effective Charts



Really the title of this slide should be "How to do Data Analysis"

1. Research

In the first step of data visualization you are finding, exploring, and understanding the limitations of your data.

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa

Who can tell me something about the dataset?

2. Edit

In the **Edit** step you are:

1. Identifying your key message
2. Choosing the best data to convey your message
3. Filtering and simplifying your data
4. Making numerical adjustments

Species	Avg Sepal Length
setosa	5.006
versicolor	5.936
virginica	6.588

What is something *else* we can explore?

Average Sepal Length and Width by Species

Let's create this together.

Species	Avg Sepal Length	Avg Sepal Width
setosa	5.006	3.428
versicolor	5.936	2.770
virginica	6.588	2.974

Difference from the Average for Each Row

Let's create this together.

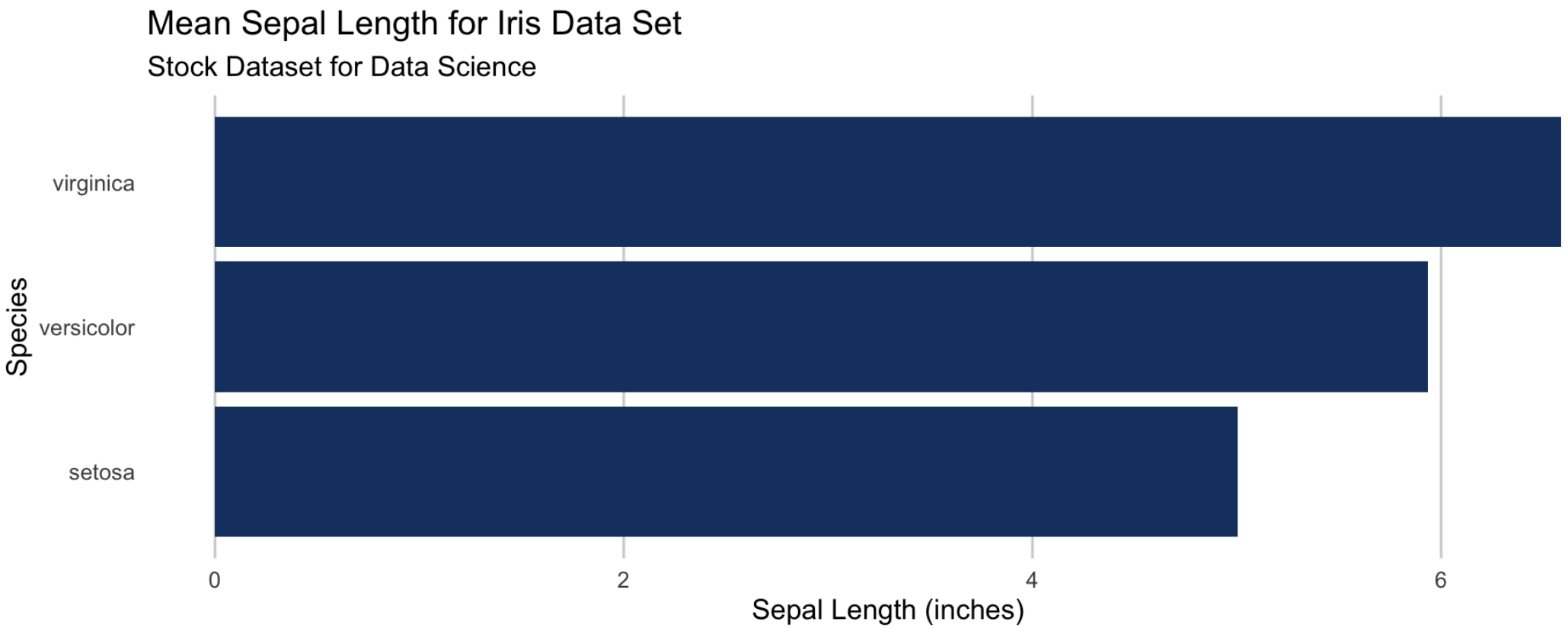
S. Len.	S. Wid.	Species	Avg. Length	Avg. Width	Pct. Diff. Len.	Pct. Diff. Wid.
5.1	3.5	setosa	5.006	3.428	0.019	0.021
4.9	3.0	setosa	5.006	3.428	-0.021	-0.125
4.7	3.2	setosa	5.006	3.428	-0.061	-0.067
4.6	3.1	setosa	5.006	3.428	-0.081	-0.096
5.0	3.6	setosa	5.006	3.428	-0.001	0.050
5.4	3.9	setosa	5.006	3.428	0.079	0.138

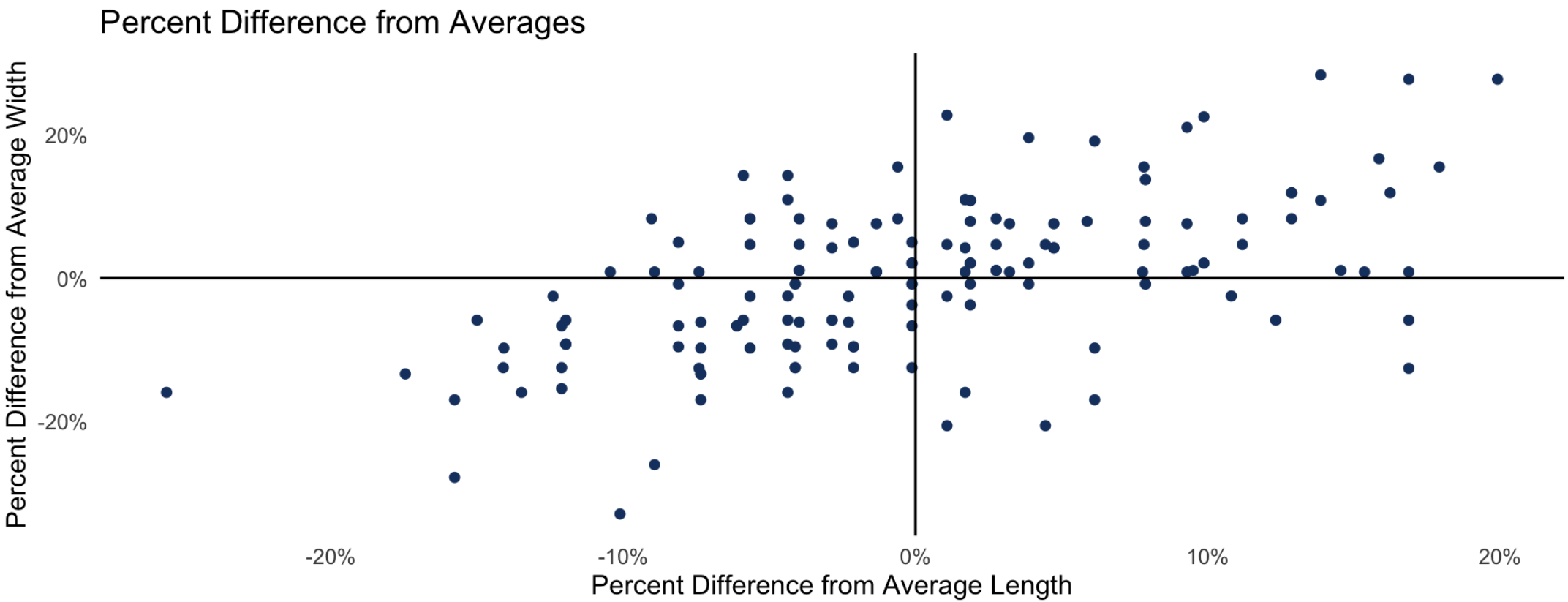
3. Plot

At this step:

1. You are choosing the proper plot type
2. Ensuring your plot axes are correct
3. Title and labels are included
4. Color and typography are clear

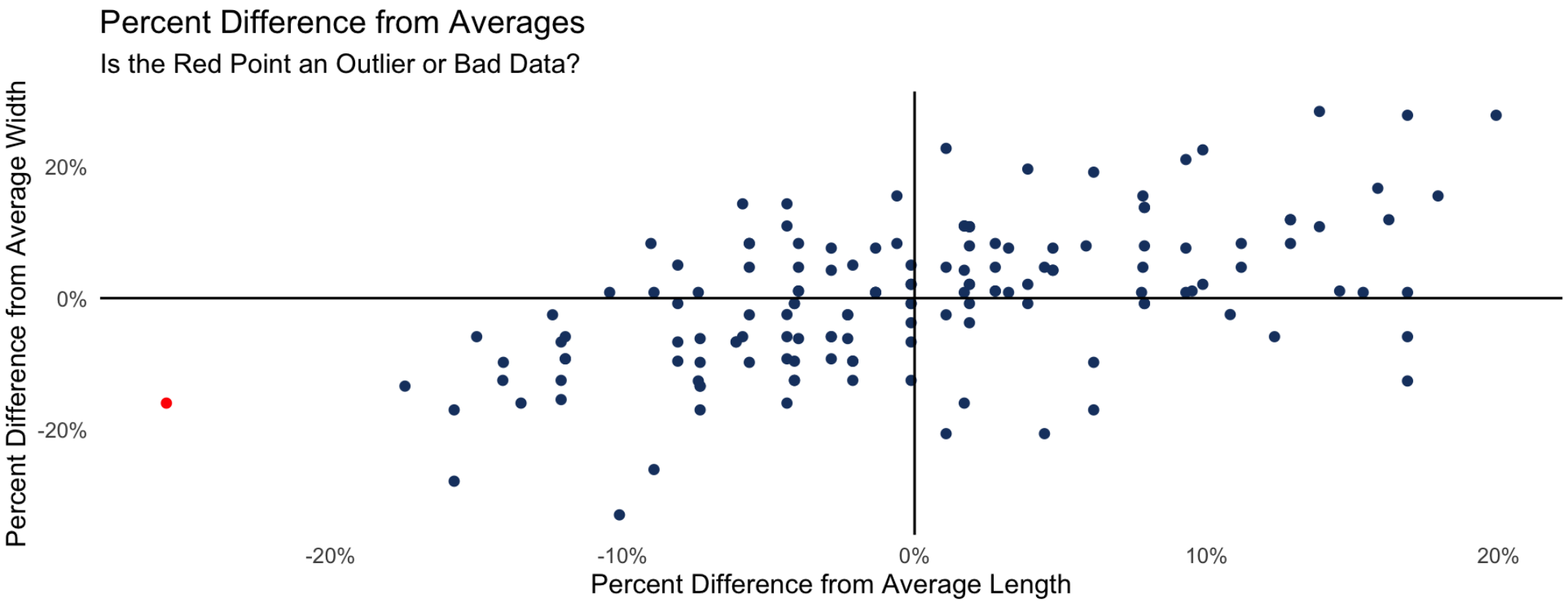
This is the reason we are here - we'll dive deeper into this shortly.


Mean Sepal Length for Iris Data Set
Stock Dataset for Data Science
Species
virginica
versicolor
setosa
0
2
4
6
Sepal Length (inches)


Percent Difference from Averages
Percent Difference from Average Width
20%
0%
-20%
-20%
-10%
0%
10%
20%
Percent Difference from Average Length

4. Review

Often overlooked due to expediency, ensuring the data is accurate will preempt questions and enhance trust in the data.


Percent Difference from Averages
Is the Red Point an Outlier or Bad Data?
Percent Difference from Average Width
20%
0%
-20%
-20%
-10%
0%
10%
20%
Percent Difference from Average Length

It is *easy* to slap something together. It is HARD to put something clean, clear, and meaningful together.

A taxonomy for data graphics

Taxonomy, smaxonomy. That's just a fancy name for saying you got to know the finer parts of a data viz before you, too, can make fine data viz.

Data graphics can be understood in terms of four basic elements:

1. Visual cues
    - We'll dive deeper into these today
2. Coordinate systems
    - We'll focus on Cartesian coordinates but there are others
    - PS: Pie charts are bad.
3. Scale
    - Logarithmic vs. Base 10
4. Context
    - Does the reader know what we are talking about?
    - Can the reader easily figure out the message?

And two bonus items:

1. Facets
    - Having multiple panels presenting similar data
2. Layers
    - Adding additional information on the original plot

Visual Cues

These are the building blocks of data viz. Visual cues are graphical elements that draw the eye to what we want our audience to focus upon.

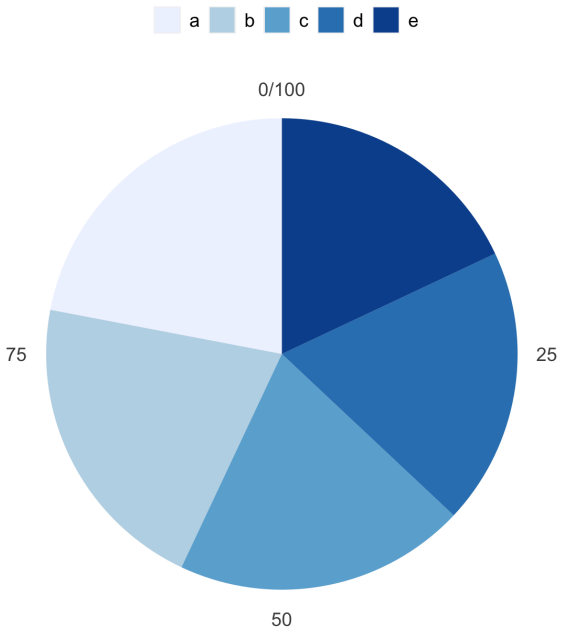
Human beings' ability to perceive difference in magnitude accurately descends in this order.

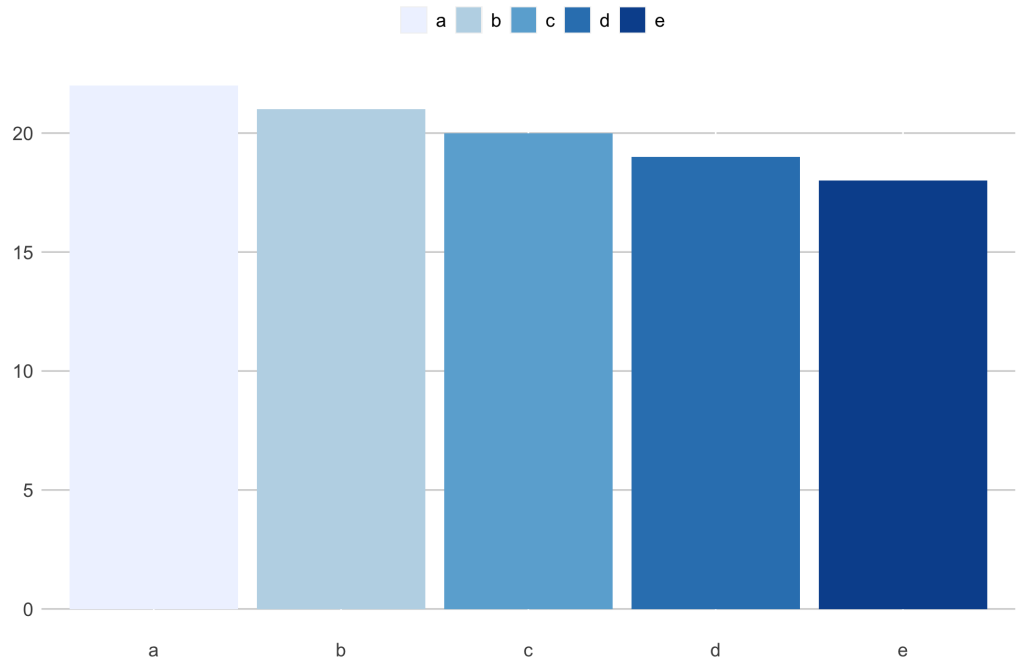
Visual cues and what they signify.

Visual Cue	Variable Type	Question
Position	numerical	where in relation to other things?
Length	numerical	how big (in one dimension)?
Angle	numerical	how wide? parallel to something else?
Direction	numerical	at what slope? in a time series, going up or down?
Shape	categorical	belonging to which group?
Area	numerical	how big (in two dimensions)?
Volume	numerical	how big (in three dimensions)?
Shade	either	to what extent? how severely?
Color	either	to what extent? how severely?

Position, Length, & Area

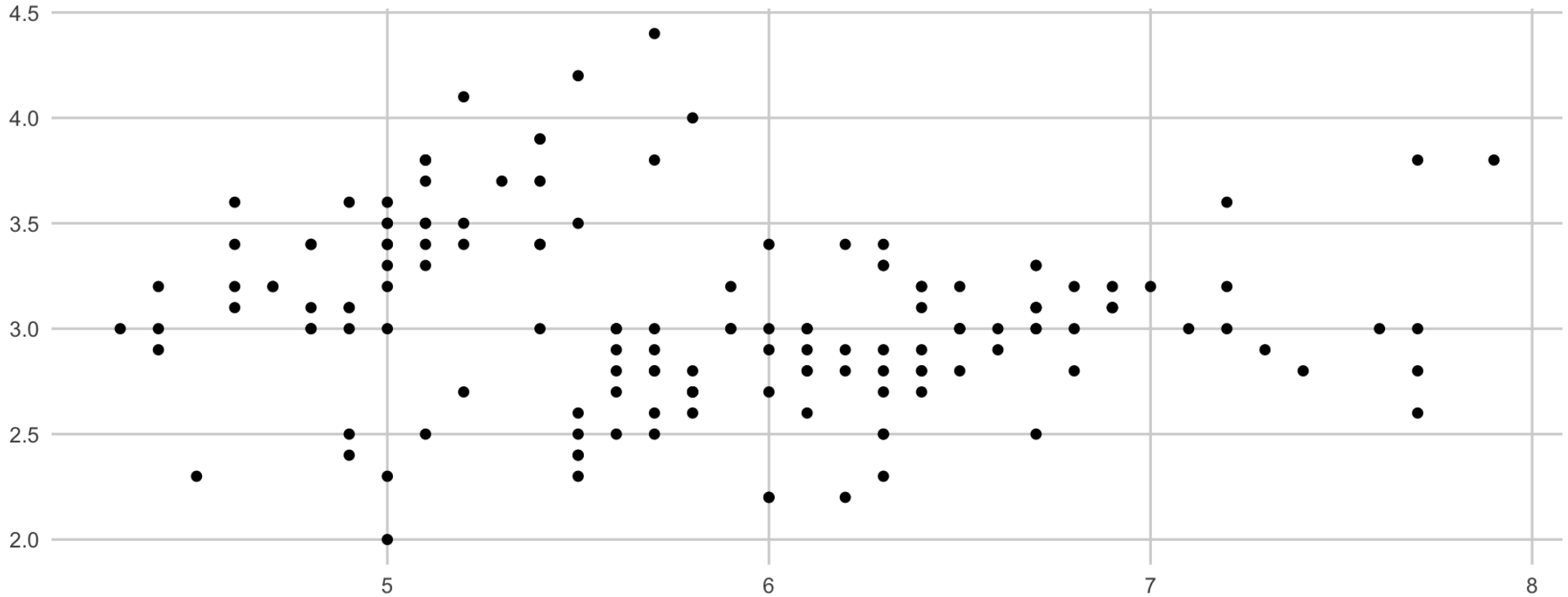
What do you see in this plot, especially as it relates to **position**, **length**, and **area**? Which letter has the most?





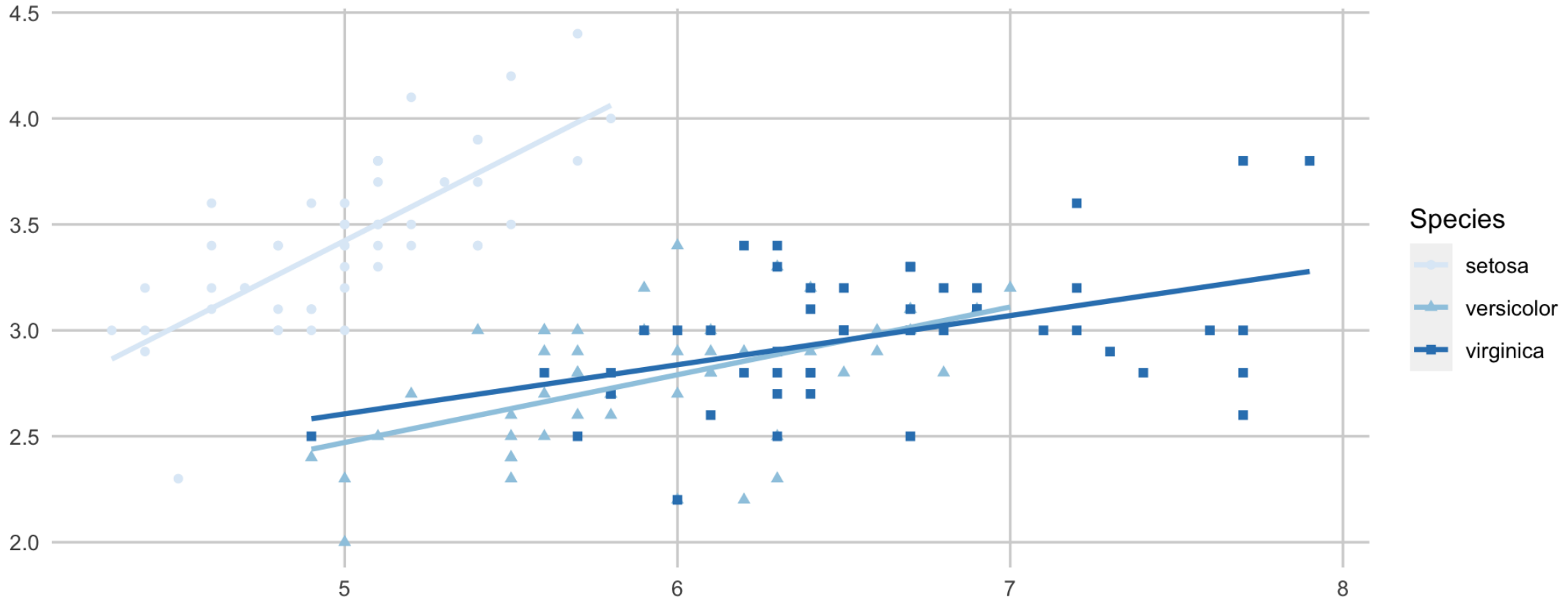
Direction

What **direction** do we see in this plot? Anything else?



Angle, Color, Shape

Talk to me about the **angle, color, and shape**.

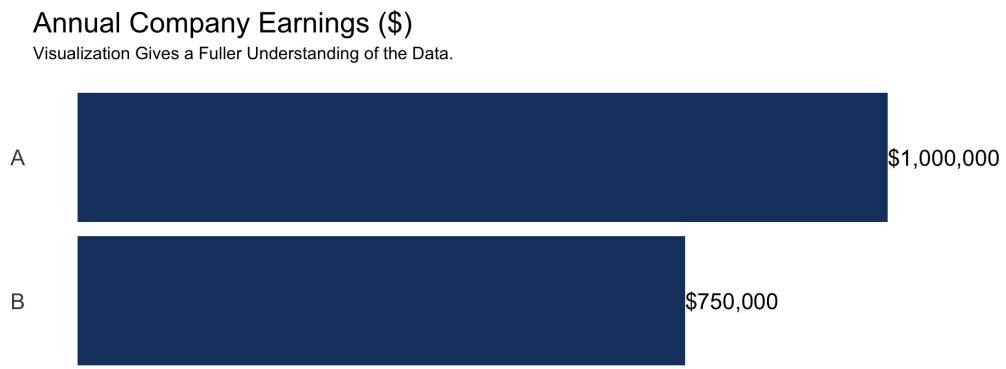


Plots Speak Louder than Words

Do NOT!

> Company A had revenues of $1MM and Company B had revenues of $750k in FY2021.

Do!



Let the Data Speak for Itself

Do not use any bs to make the graph "pretty." A clean crisp chart allows the reader to focus on the data and the message of the story.

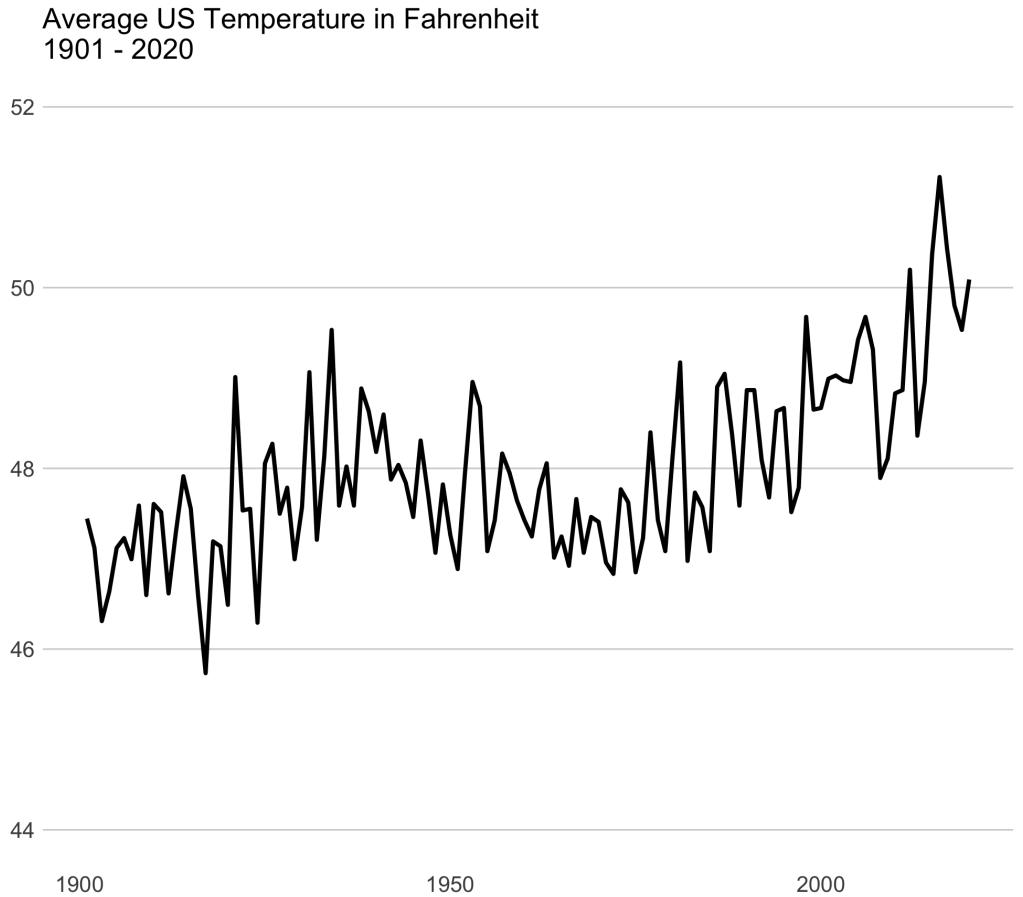
This means:

1. No 3D rendering EVER
2. No "pretty" colors
3. No heavy grid lines
4. NEVER NEVER NEVER a pie chart (though the WSJ doesn't agree with me)

Misrepresenting the Trend

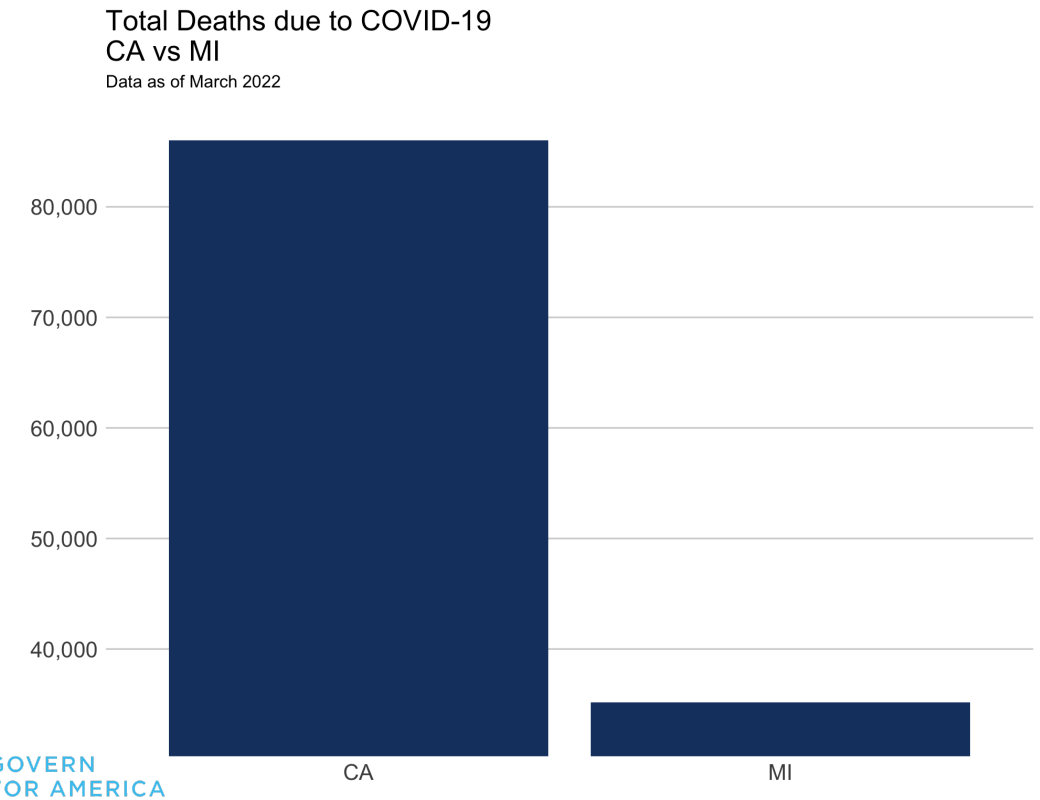
Misrepresenting the Trend

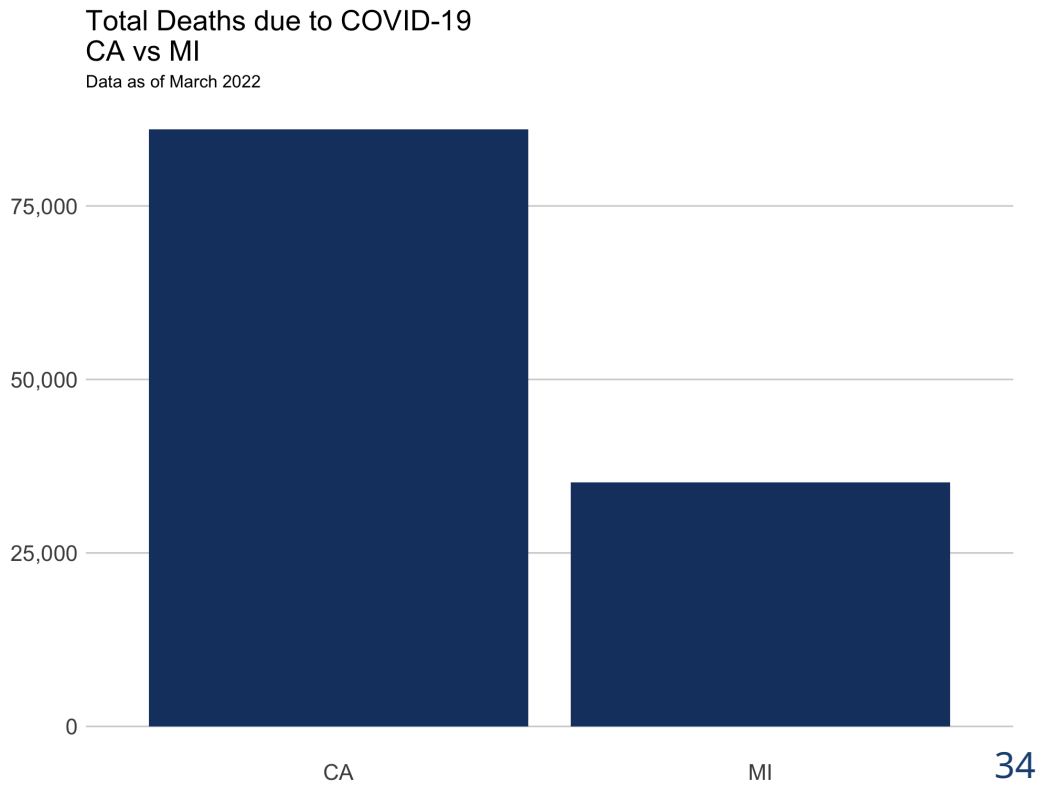
- The data should fill about 2/3 of the y-axis height
- The line should not be too thick or thin
- The grid shouldn't be thicker than the line
- The y-axis increments shouldn't be awkward
  - Natural increments: (1s, 2s, 5s, 10s, etc.)
  - Awkward increments: (3s, 6s, 8s, 12s, etc.)



Zero Baseline

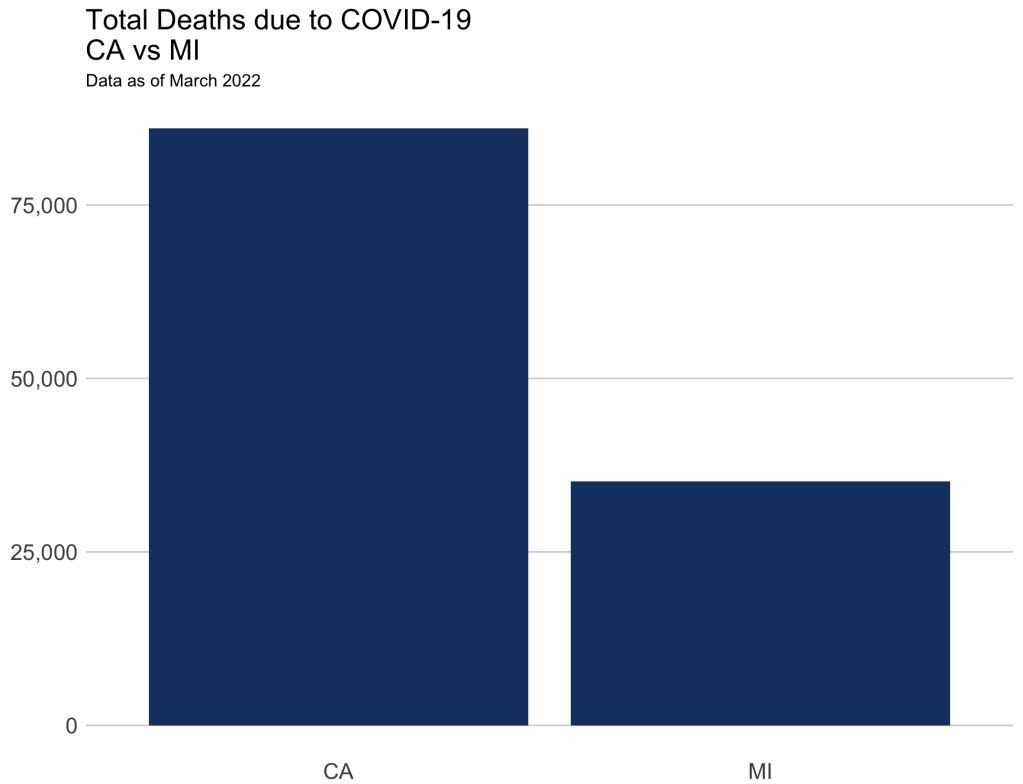
- With line charts, as we just saw, we want the y-axis to be the right size
- With bar charts, we are interpreting area; because of this, we always start at 0





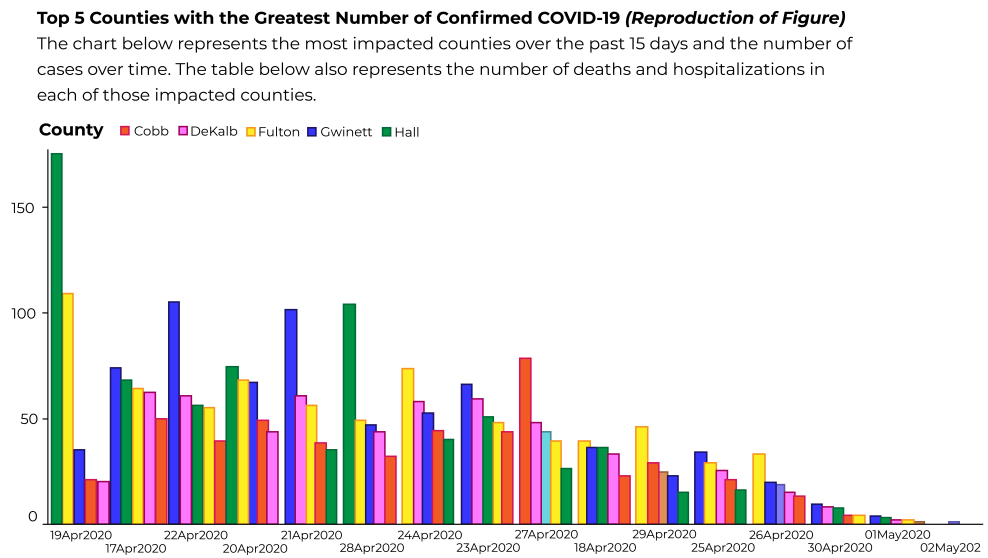
Zero Baseline

- We can clearly see that MI has about half the total COVID Deaths as CA
- Even this plot may still be misleading or hard to interpret: CA and MI have vastly different populations (should probably be a percentage of population)



Order Bars in the Correct Order

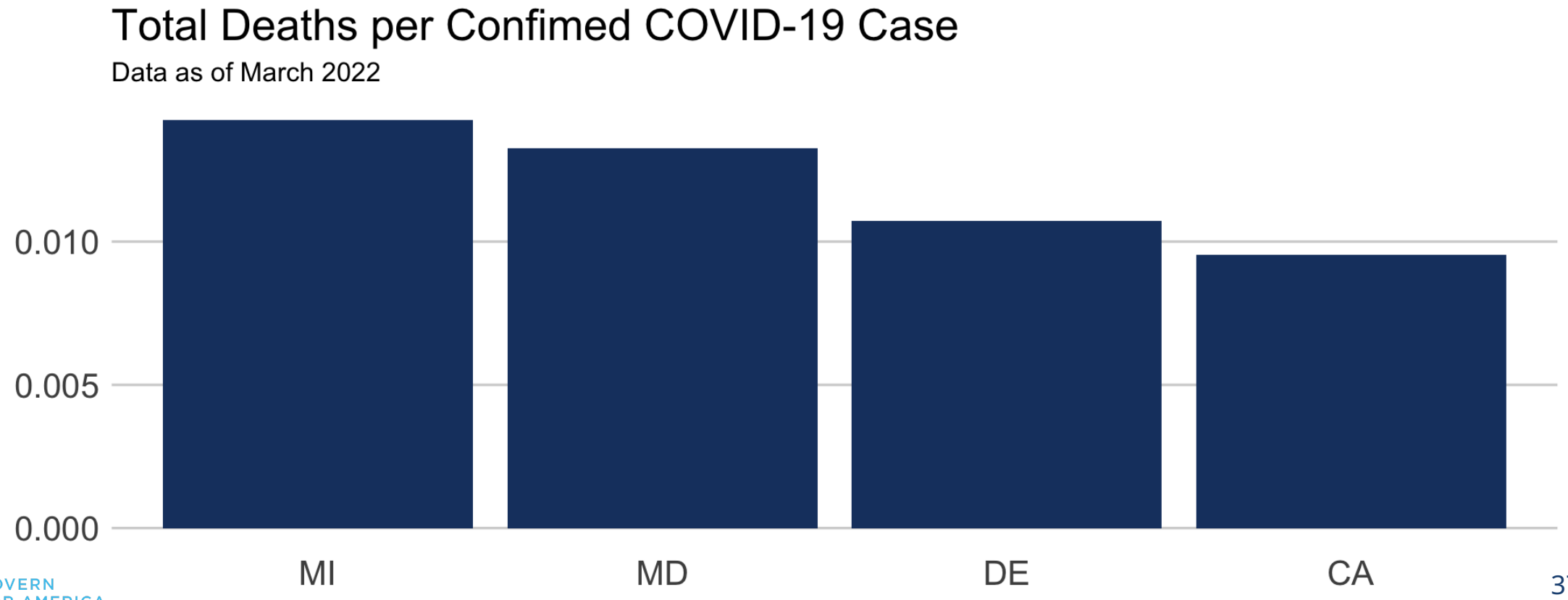
This is not good (look at the dates).



- Ordinal data should stay in order
  - Grades
  - Dates
  - Age groups (Child, Teen, Young Adult, Adult, etc)

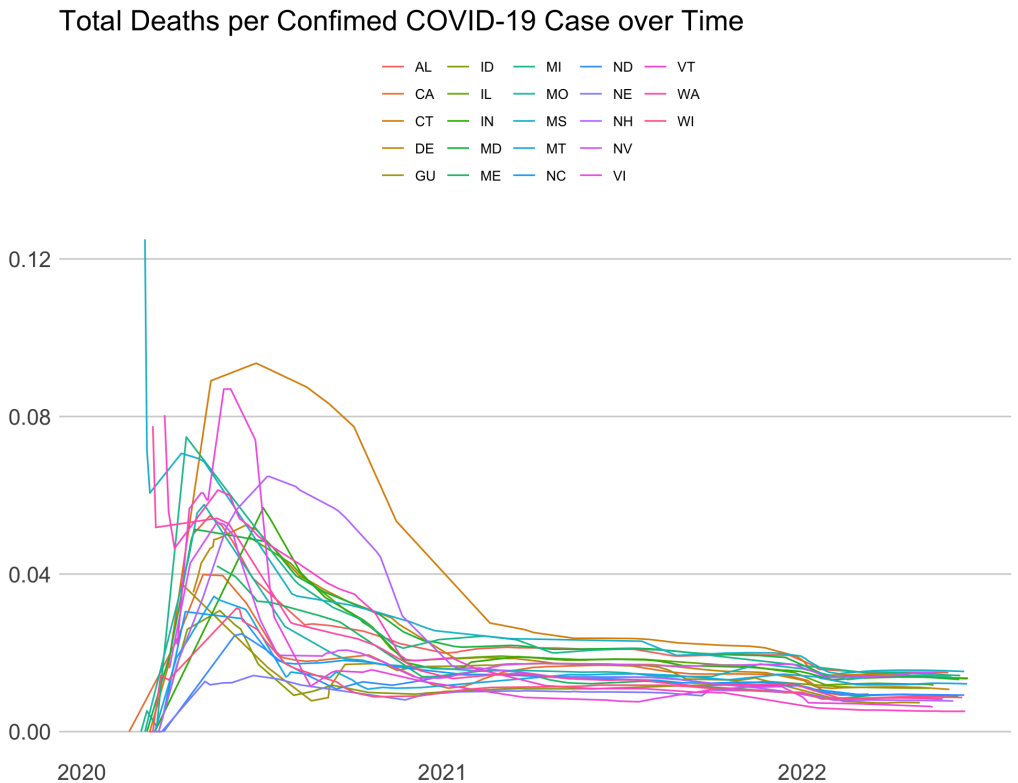
Order Bars in the Correct Order

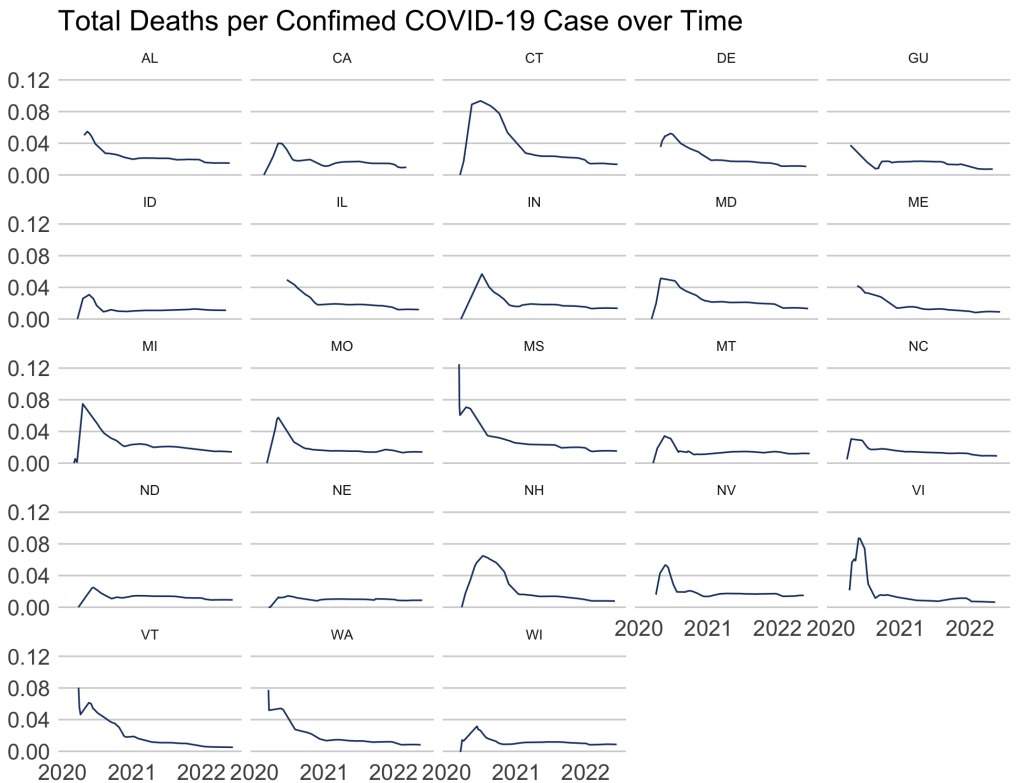
Other categorical data should be reordered so the "highest" is on the left and "lowest" is on the right.



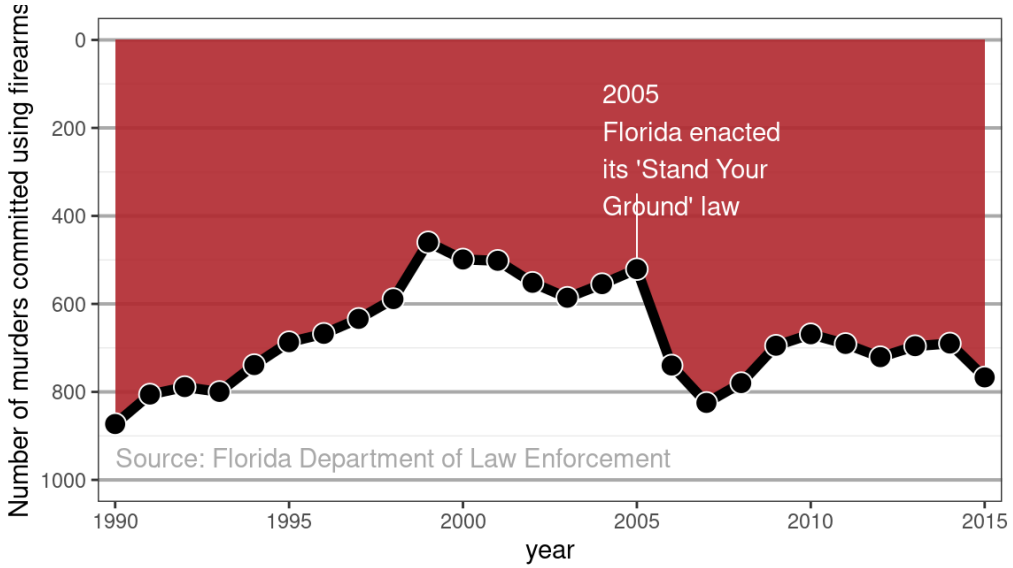
Too many lines on a plot

When you have too many categories, you should split your plot into separate "facets."





Negative Values



This is another example of what not to do.

Any values under the x-axis or left of the y-axis are treated naturally as negative values. It is misleading to flip the axes.

Another Exercise Together!

https://tinyurl.com/GFA-Final-Exercise

Learning More

If you got to this slide and are thinking to yourself... "I'm good on knowing Data Visualization now," then this presentation is a good baseline for you.

To understand a bit more, buy and read Guide to Information Graphics

Learning More

If, on the other hand, you are feeling an itchiness inside to learn more and more and more - AWESOME. Welcome to your journey.

"If I don't get into this nerd school, I'm going to lose my mind" - Hiro Hamada

Learning More

1. Tableau or other Data Visualization Tool
    - Tableau is the defacto "business intelligence" tool that companies use
    - It is super powerful for creating many types of data visualizations in many formats
2. SQL
    - Tableau, however powerful it may be, can only visualize the data which it has access
    - In many cases, data is stored in databases and access via a language called SQL
3. Data analysis in R
    - If you start creating SQL code and want to do more, R may be the next best languague to learn
    - Harvard has some amazing, free, resources on edx to learn data and computer science.

Thank You!

If you have any questions as you go through your fellowship and beyond, please reach out!