Introduction to Data Visualization

STAT 4380 - 001

Jake Rozran

Villanova University

2022-02-01

The 2012 Federal Election Cycle

To Begin, a Case Study!

We are going to use an example from the book to show how important choices in data visualization effect our understanding of what is happening.

Ethics Note: When done intentionally, this is how one may "lie with statistics."

# The Federal Election Commission (FEC) Maintains a Website of All Contributions

On this website there are:

1. All contributions of $200 or more made by individuals to candidates and committees.
2. Spending by committees on behalf (and against) candidates.
3. In fact, this data is available in R in the `fec12` library!

# If you had this Data and Were Tasked with an Analysis, What Questions Might you Ask?

A Good Place to Start: "On whose candidacy was the most money spent?"

What Conclusions do you Draw From the Data?

Money Spent (millions of USD)
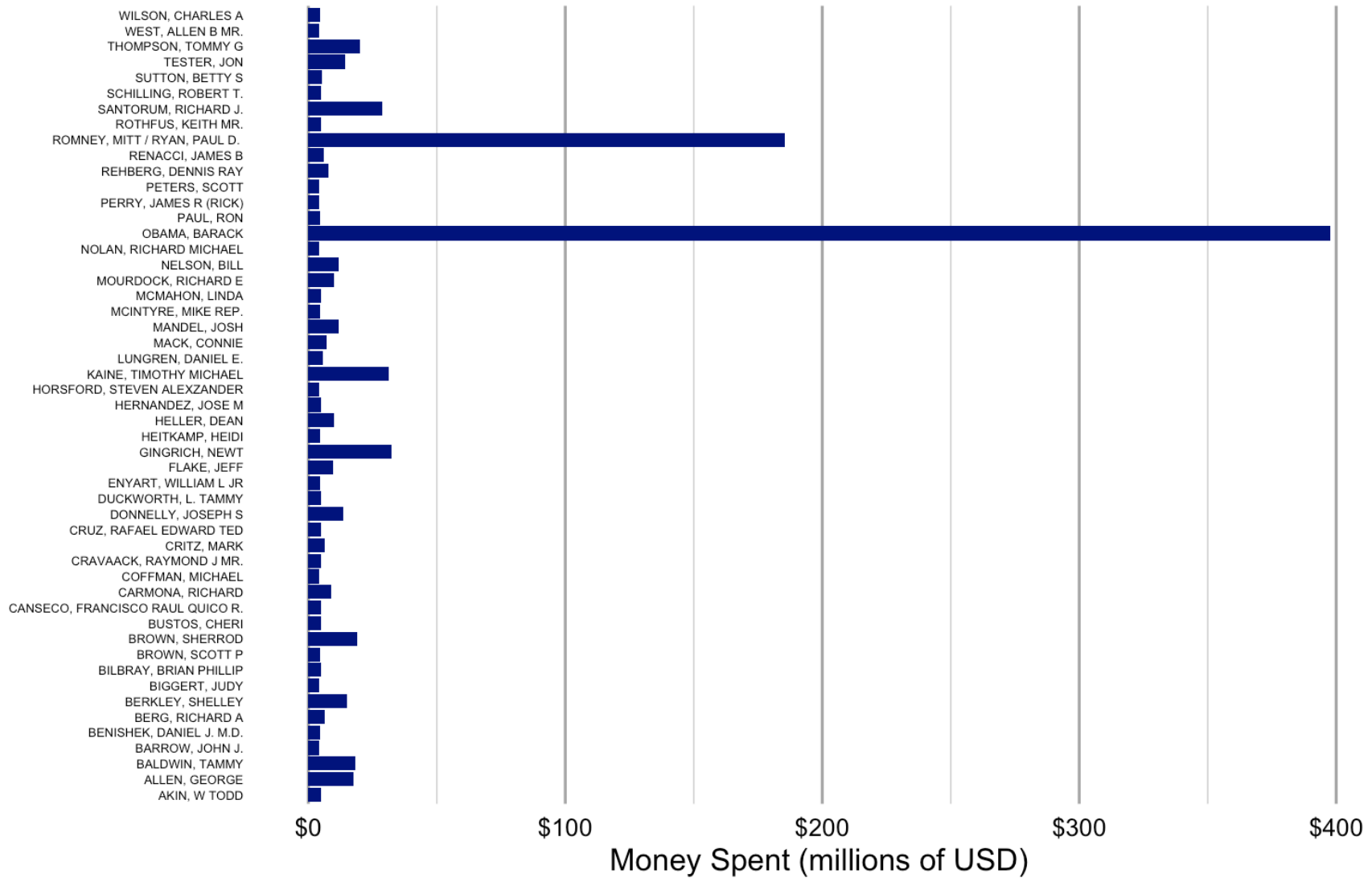
Fun (and Obvious Fact): Committees are not Limited to Spending Money in Support of a Candidate

They Can also Spend Money **against** a candidate. Now what conclusions do you draw? This is the same exact data as before, with more information provided.
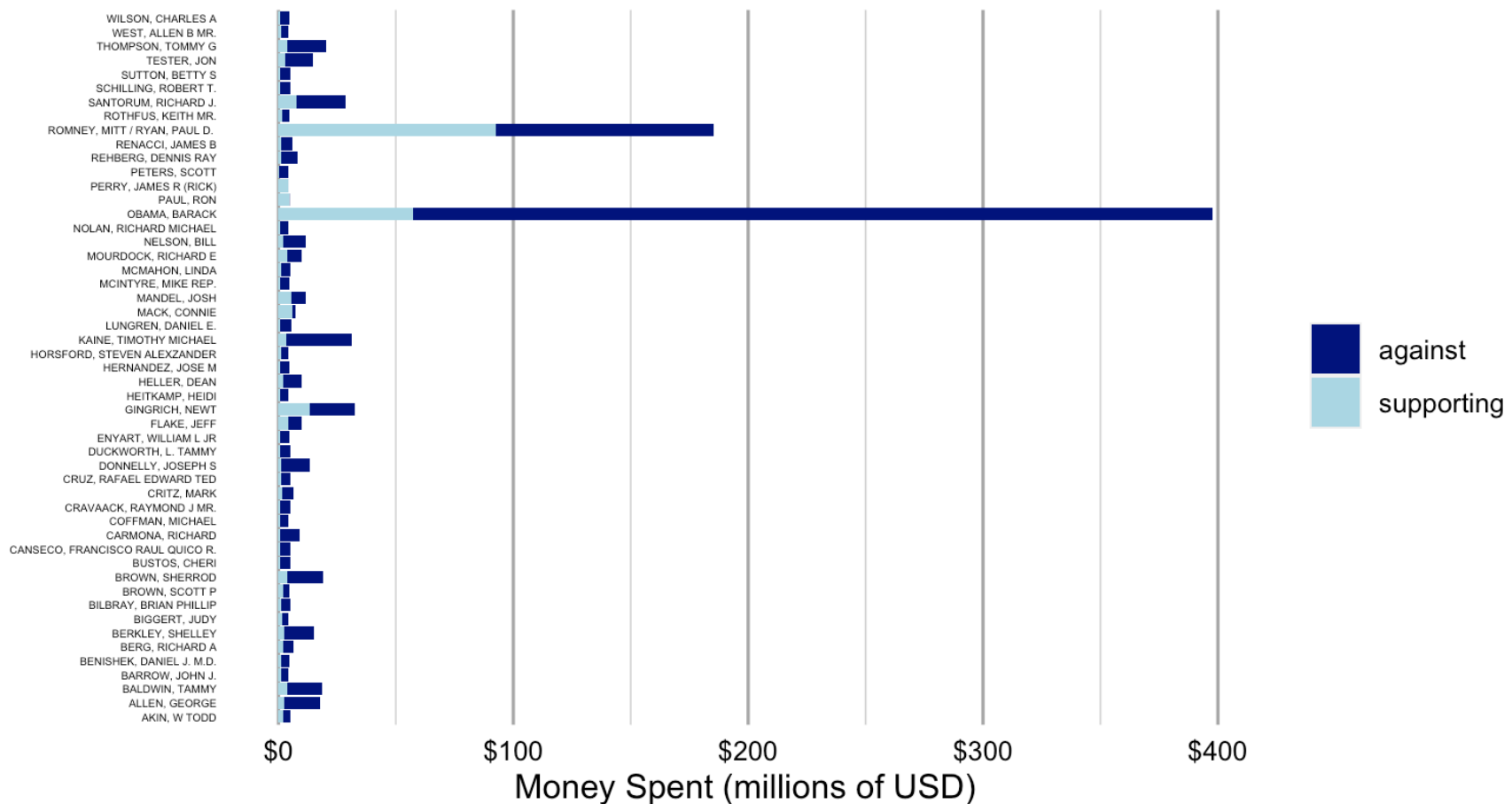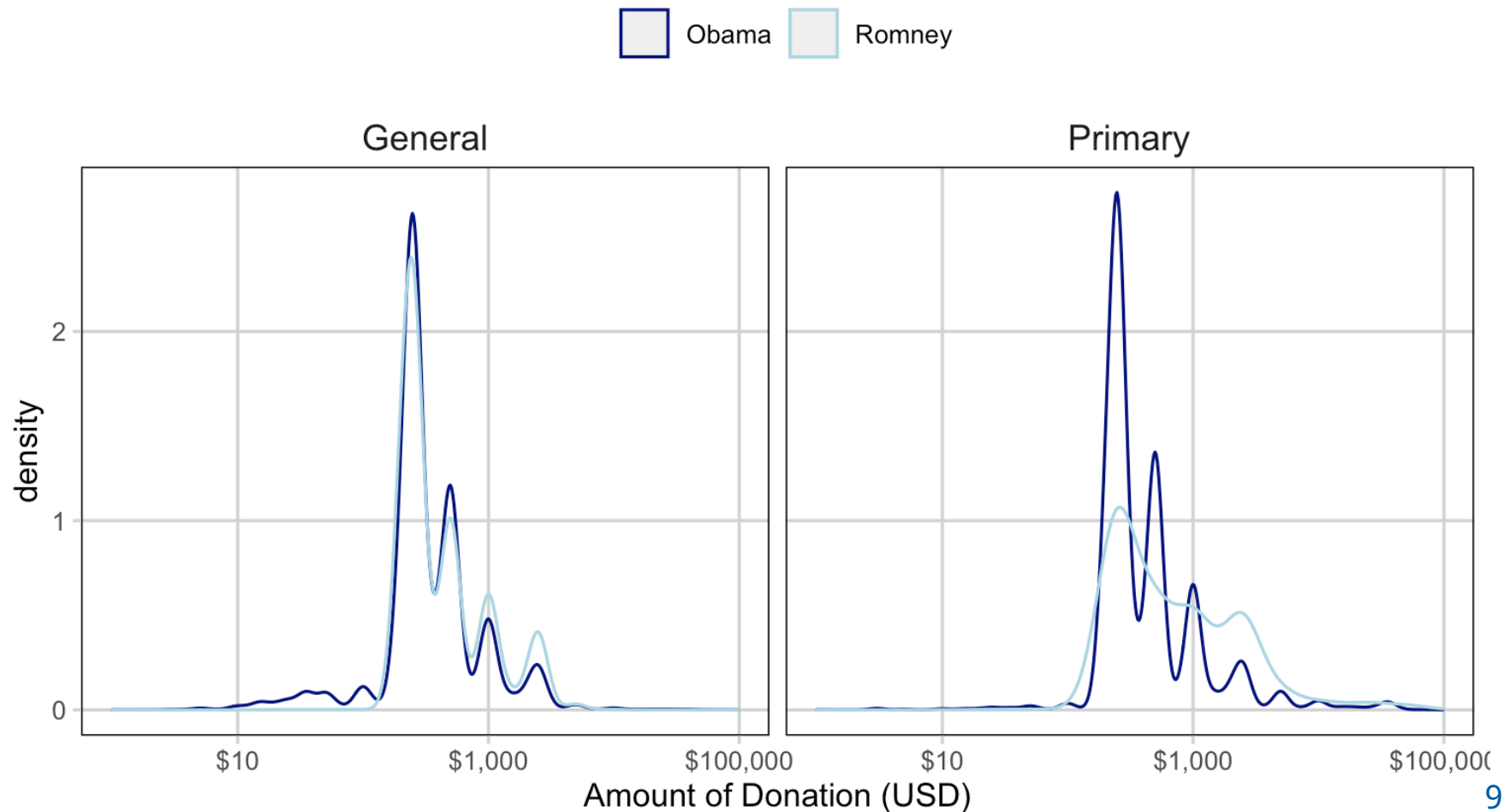
Another Question to Explore: Was Mitt Supported by a Few Rich Donors While Barack Supported "by the People"?

- If we look at the whole election, what conclusion would you draw?
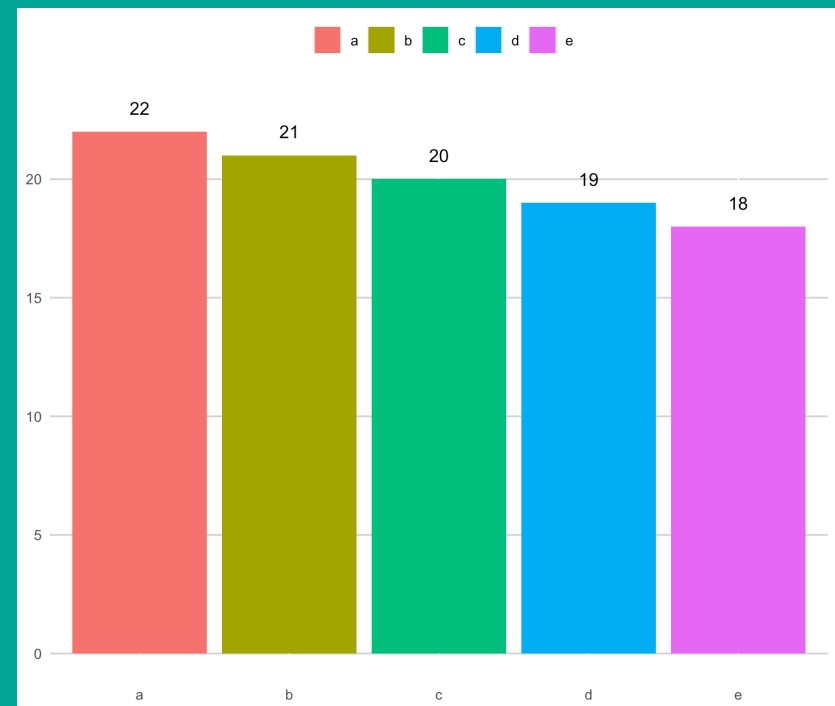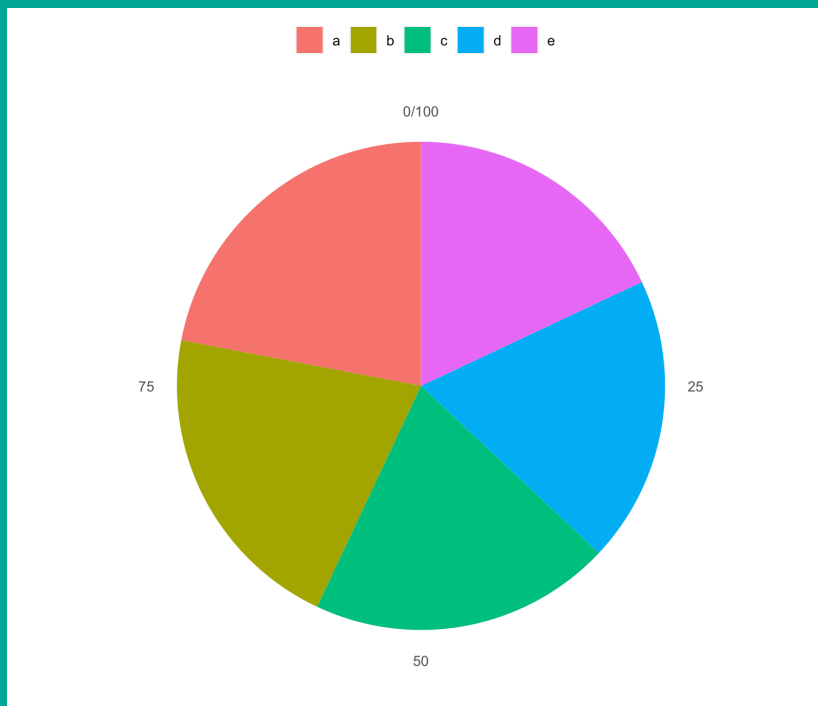- What else might we ask of this data?

Diving Deeper: Let's Change the Data and Again Answer the Same Question

Was Mitt Supported by a Few Rich Donors While Barack Supported "by the People"?

Composing Data Graphics

Anyone can create a data visualization - the hard part is making it tell the correct story in the easiest to understand way.

To the Point of the Last Slide, it has Taken me Hours to the Previous Slides Just how I Want Them

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
2. Coordinate systems
3. Scale
4. Context

And two bonus items:

1. Facets
2. Layers

Visual Cues

- Visual cues are graphical elements that draw the eye to what we want our audience to focus upon
- These are the building blocks of data viz
- Human beings' ability to perceive difference in magnitude accurately decends in this order
  - For example, humans are not as good at perceiving angles as we are length, hence why the example before with the pie chart was so difficult
  - Also, humans are poor at perceiving differences in color, that is why being thoughtful about color is so important (and why heatmaps are frequently poopooed)

Visual Cues (Continued)

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 Example

Because I feel like this is getting a little too much me talking at you, let us do an example together.

Someone tell me what they see in this plot, especially as it relates to **position, length, and area**?
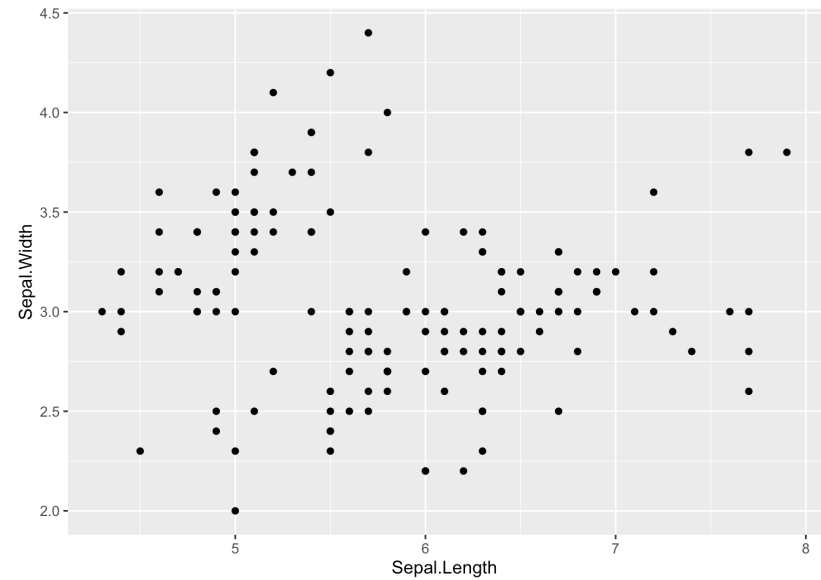
```
plot_dat <- iris %>%
    group_by(Species) %>%
    summarise(avg_sepal_length = mean(Sepal.Length))

ggplot(plot_dat, aes(x = Species, y = avg_sepal_length)) +
    geom_bar(stat = "identity")
```
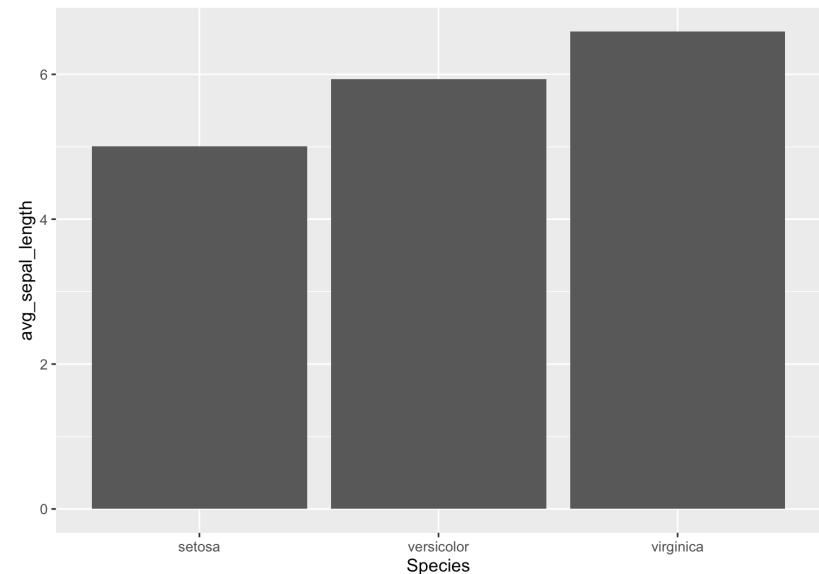
Direction

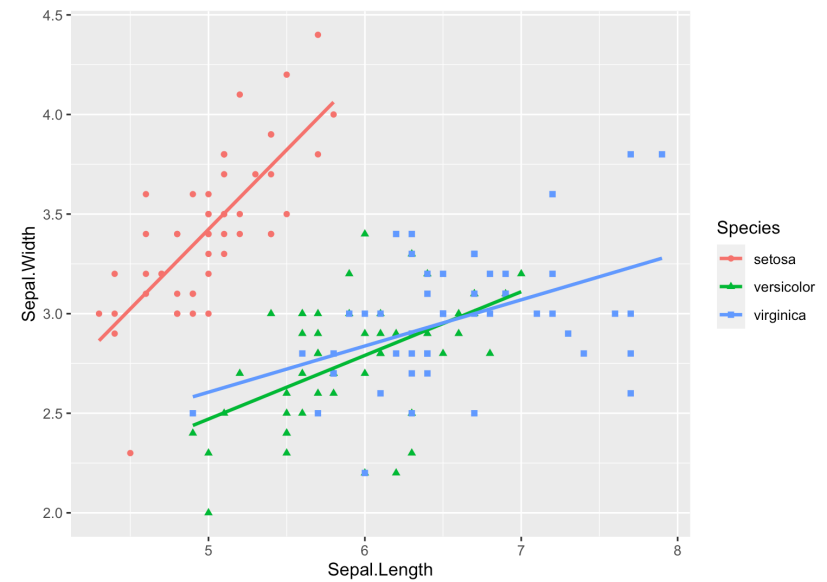What **direction** do we see in this plot? Anything else?

```
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
    geom_point()
```

Angle, Color, Shape

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

```
ggplot(iris, aes(x = Sepal.Length,
                 y = Sepal.Width,
                 color = Species,
                 shape = Species)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE)
```

Angle, Color, Shape and Size

We've just added another dimension: **size**. What do we realize now?

```
ggplot(iris, aes(x = Sepal.Length,
                 y = Sepal.Width,
                 color = Species,
                 shape = Species,
                 size = Petal.Length)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE)
```