Data Visualization

Data Science Collaborative

Ethan P. Marzban

Department of Statistics and Applied Probability; UCSB

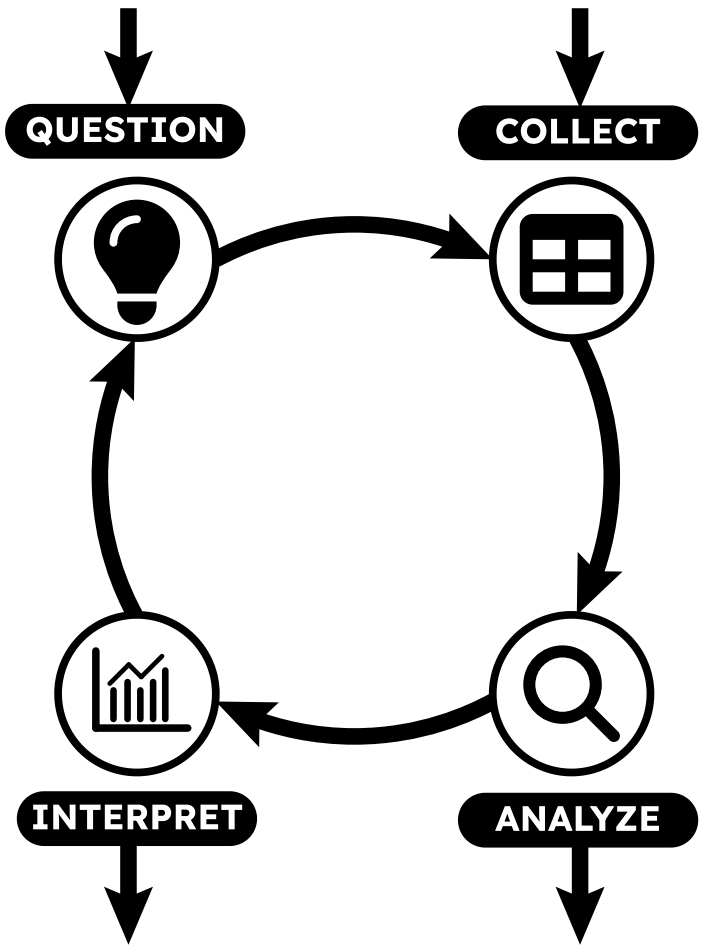
November 5, 2025

? Some Questions

Leadup

- How do the GDPs of countries vary as a function of average life expectancy at birth?
- Does the nature of this relationship change across continents?
- Can you *justify* your answers?

  → What sort of **data** could be used to answer these questions *and* provide appropriate justification?



World Bank Dataset

- The *World Bank* is a collection of organization aiming to study the effects of poverty worldwide.
  - → You can read more about them at their **website**.
- Some variables with corresponding data:

- Country Name
- Country Code (abbreviation)
- Continent
- Year of observation
- GDP (Gross Domestic Product)
- Female Life Expectancy at Birth
- Male Life Expectancy at Birth
- Total Life Expectancy at Birth
- Female Adult Literacy Rate
- Male Adult Literacy Rate
- Total Adult Literacy Rate
- Female Youth Literacy Rate
- Male Youth Literacy Rate
- Total Youth Literacy Rate
- Population

World Bank Dataset

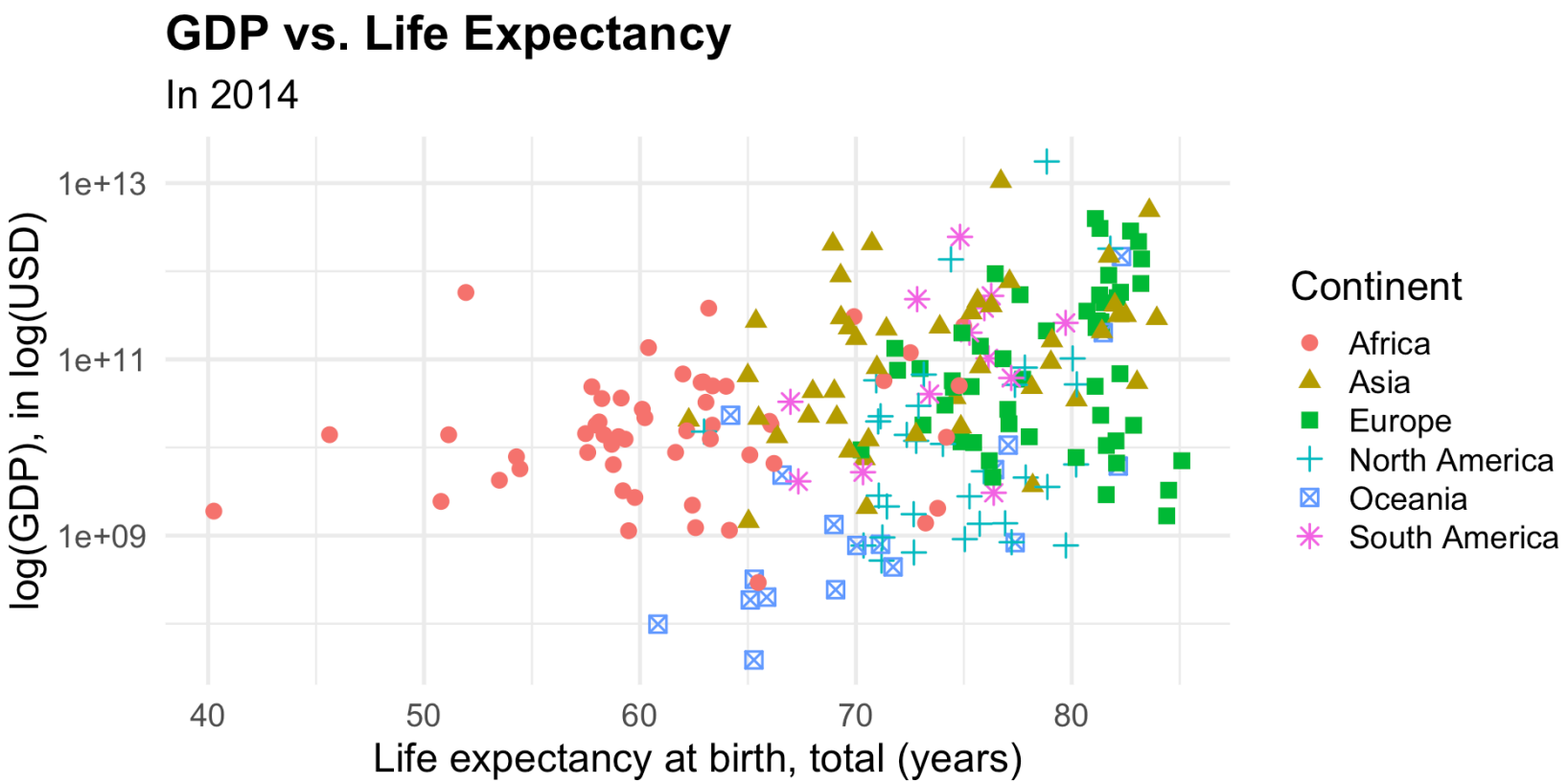
- How do the GDPs of countries vary as a function of average life expectancy at birth?
- Does the nature of this relationship change across continents?

```r
wb <- read.csv("data/wb_cont.csv", check.names = FALSE)
wb %>% head(100)
```

```
          Country Name Country Code     Continent
1            Afghanistan          AFG          Asia
2            Afghanistan          AFG          Asia
3            Afghanistan          AFG          Asia
4            Afghanistan          AFG          Asia
5            Afghanistan          AFG          Asia
6            Afghanistan          AFG          Asia
7            Afghanistan          AFG          Asia
8            Afghanistan          AFG          Asia
9            Afghanistan          AFG          Asia
10           Afghanistan          AFG          Asia
```

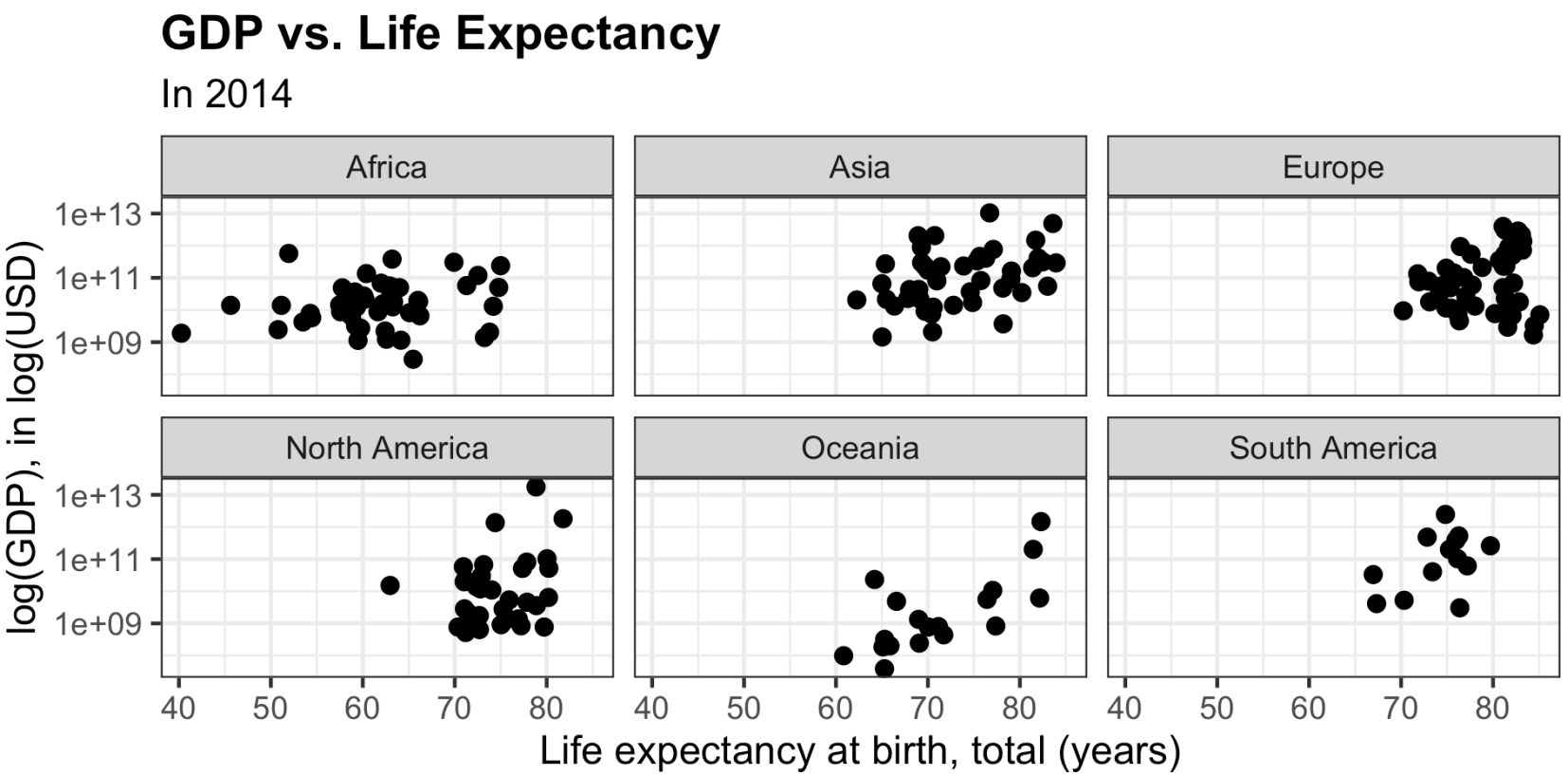
World Bank Dataset

- How do the GDPs of countries vary as a function of average life expectancy at birth?
- Does the nature of this relationship change across continents?



World Bank Dataset

- How do the GDPs of countries vary as a function of average life expectancy at birth?
- Does the nature of this relationship change across continents?



Data Visualizations

Overview

- Notice how, with just a single well-crafted visualization, we were able to answer our initial questions with ease!
- This illustrates one of the major reasons why visualizations are so important: they can succinctly summarize data, and highlight important patterns that would be otherwise very difficult (or impossible) to see.
- My goal in this workshop is to help you craft **presentation-quality** graphics, which are highly curated for maximal impact.

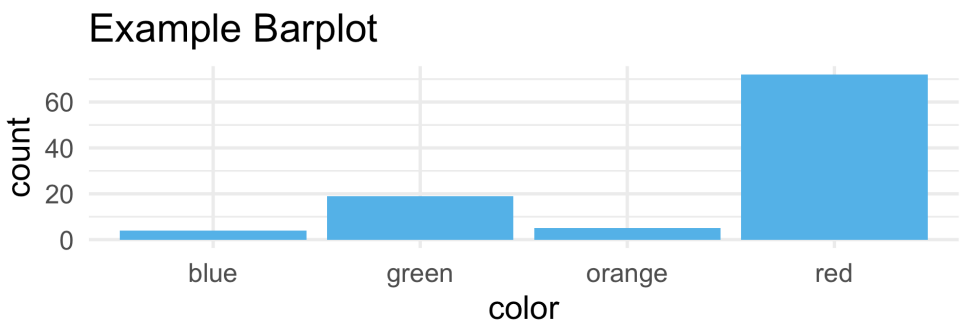
  → Contrast these with **exploratory** visualizations, which are more "quantity over quality".

Data Visualizations

Basic Building Blocks

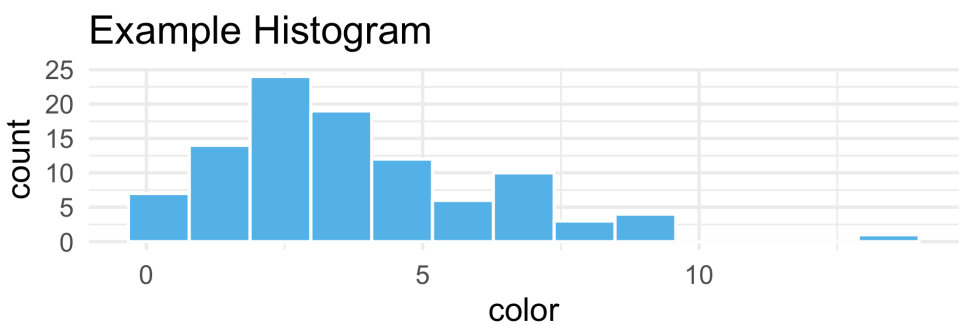
Univariate Categorical Data:

- **Bargraphs** (aka **barplots**)



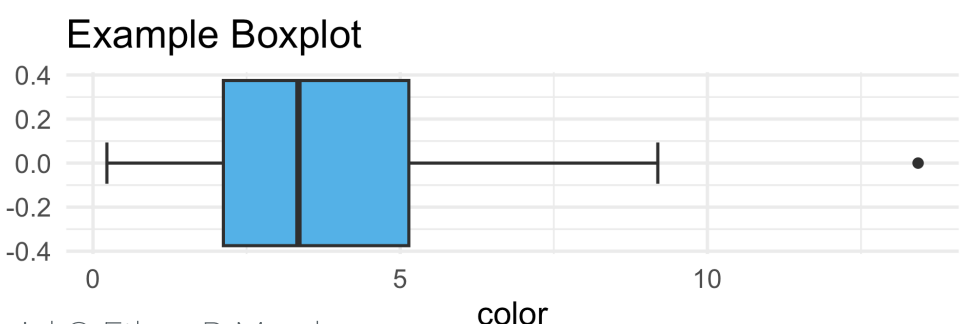
Univariate Numerical Data:

- **Histograms**
- **Boxplots**



Bivariate Numerical Data:

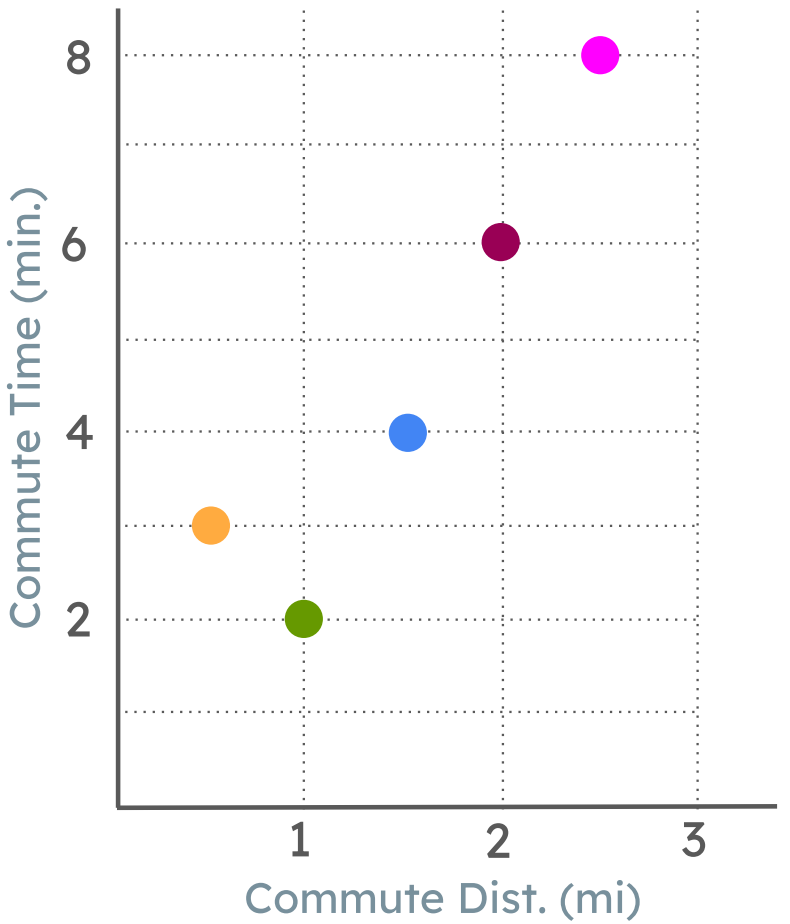
- **Scatterplot**



Data Visualizations

Scatterplots

Commute Dist.	Commute Time
0.5	3
1	2
1.5	4
2	6
2.5	8



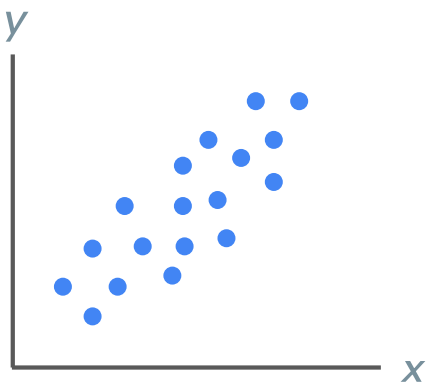
Scatterplot

Trends

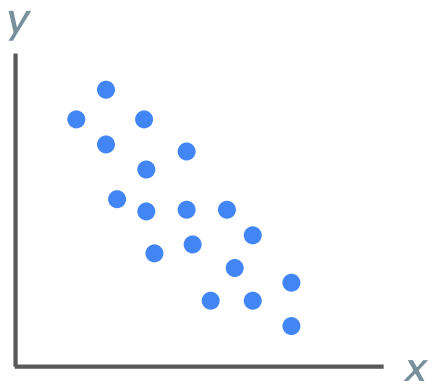
- When considering scatterplots, certain patterns may become apparent.
  - → For example, notice that, on average, as commute distance increases, so does commute time.
- Such patterns are called **trends**.
- Most trends can be classified along two axes: positive/negative, and linear/nonlinear.
- A **positive** trend is observed when as x increases so does y; a **negative** trend is observed when as x increases y decreases.
- A trend whose rate of change is constant is said to be **linear**; a trend whose rate of change is nonconstant is said to be **nonlinear**

Scatterplot

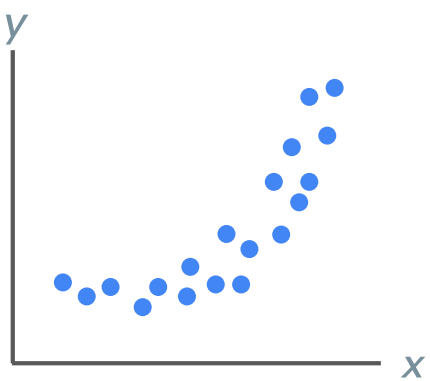
Trends



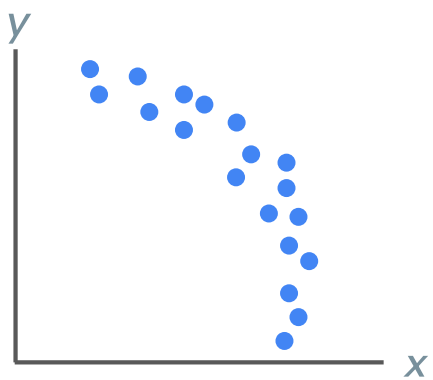
- **Positive**: as x increases, y also increases
- **Linear**: the rate of change is roughly constant



- **Negative**: as x increases, y decreases
- **Linear**: the rate of change is roughly constant



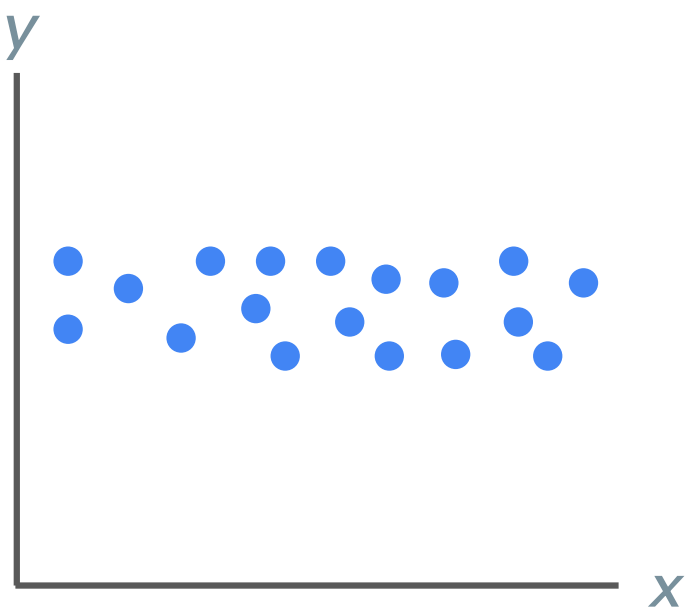
- **Positive**: as x increases, y also increases
- **Nonlinear**: the rate of change is nonconstant



- **Negative**: as x increases, y decreases
- **Nonlinear**: the rate of change is nonconstant

Scatterplot

Trends



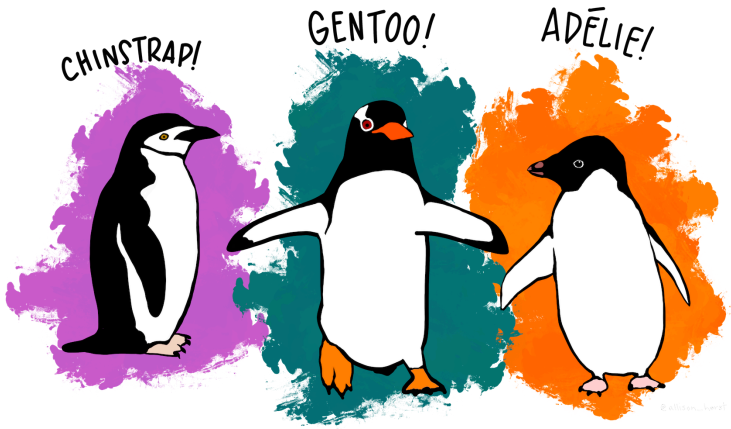
- **No trend.**
  - **As *x* increases, *y* appears to stay relatively constant.**

- Another way to describe the findings of a scatterplot is in terms of the **association** between the variables being compared.

  → For instance, if the scatterplot of y vs. x displays a positive linear trend, we would say that x and y have a positive linear association, or that x and y are positively linearly associated.

Penguins

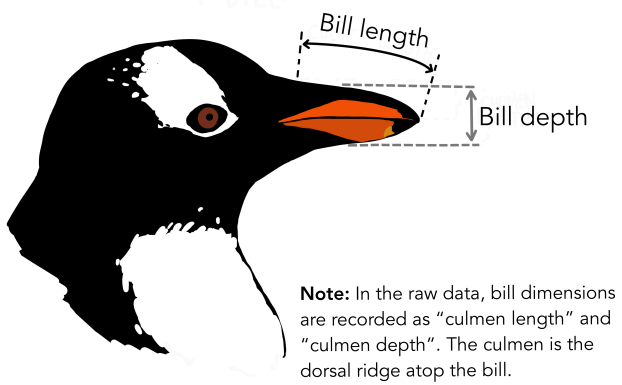
An Example



Artwork by @allison_horst

- The `penguins` dataset, from the `palmerpenguins` package, contains information on 344 penguins, collected by Dr. Kristen Gorman, at the Palmer Research Station in Antarctica.
- Three species of penguins were observed: Adélie, Chinstrap, and Gentoo

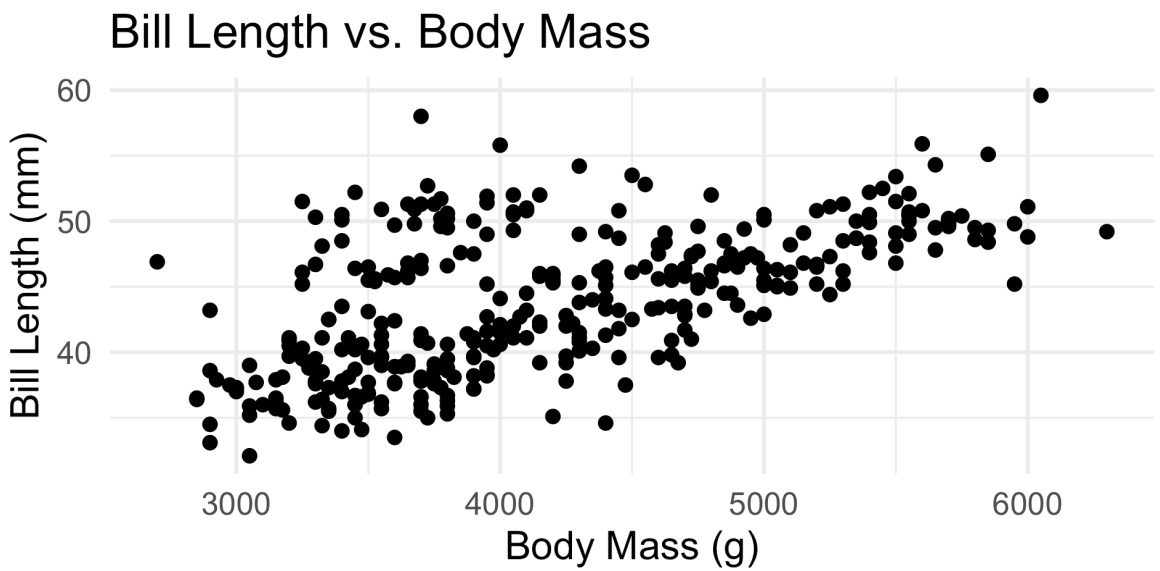
- Various characteristics of each penguin were also observed, including: flipper length, bill length, bill depth, sex, and island.
- It seems plausible that a penguin's bill length should be related to its body mass.



Artwork by @allison_horst

Penguins

An Example



- Is there a trend?
  - → Increasing or decreasing?
  - → Linear or nonlinear?
- Do heavier penguins tend to have longer bills?

Question

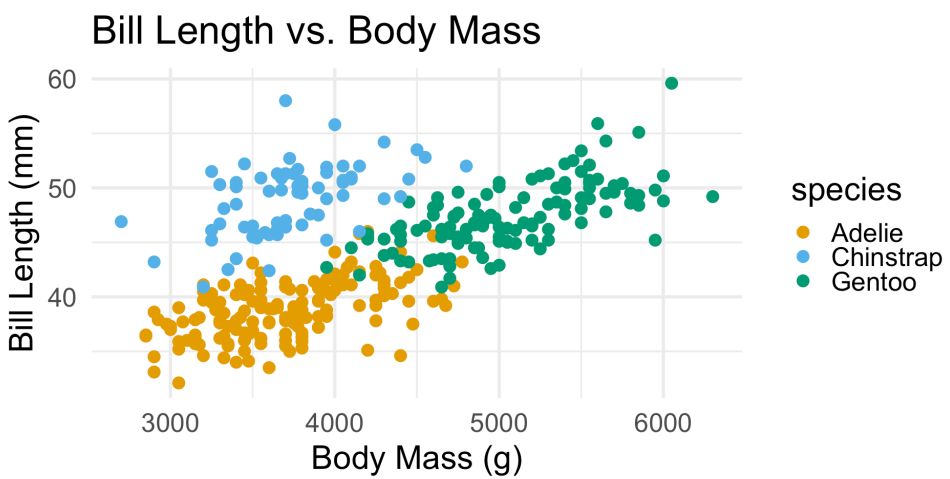
Does our answer change depending on the penguins' species?

- That is, do different species exhibit different relationships between body mass and bill length?

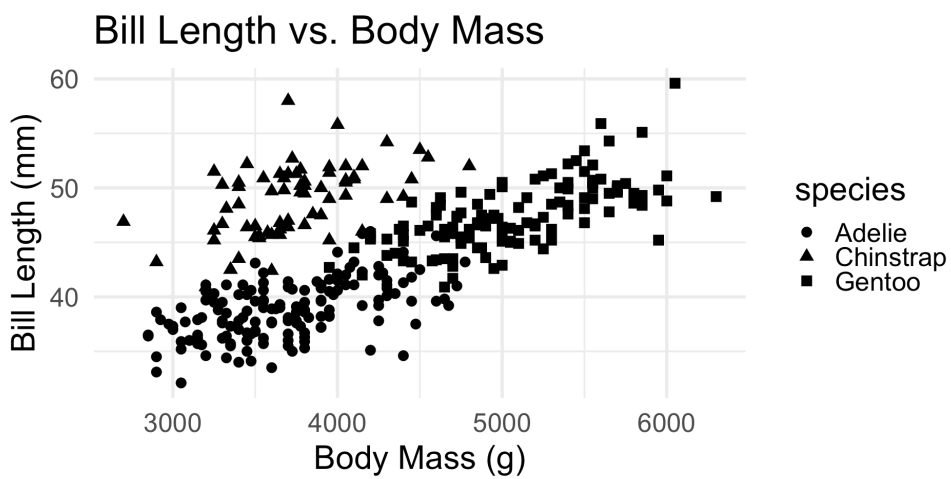
Penguins

Two Ideas:

First Idea: Color each point according to the associated species:



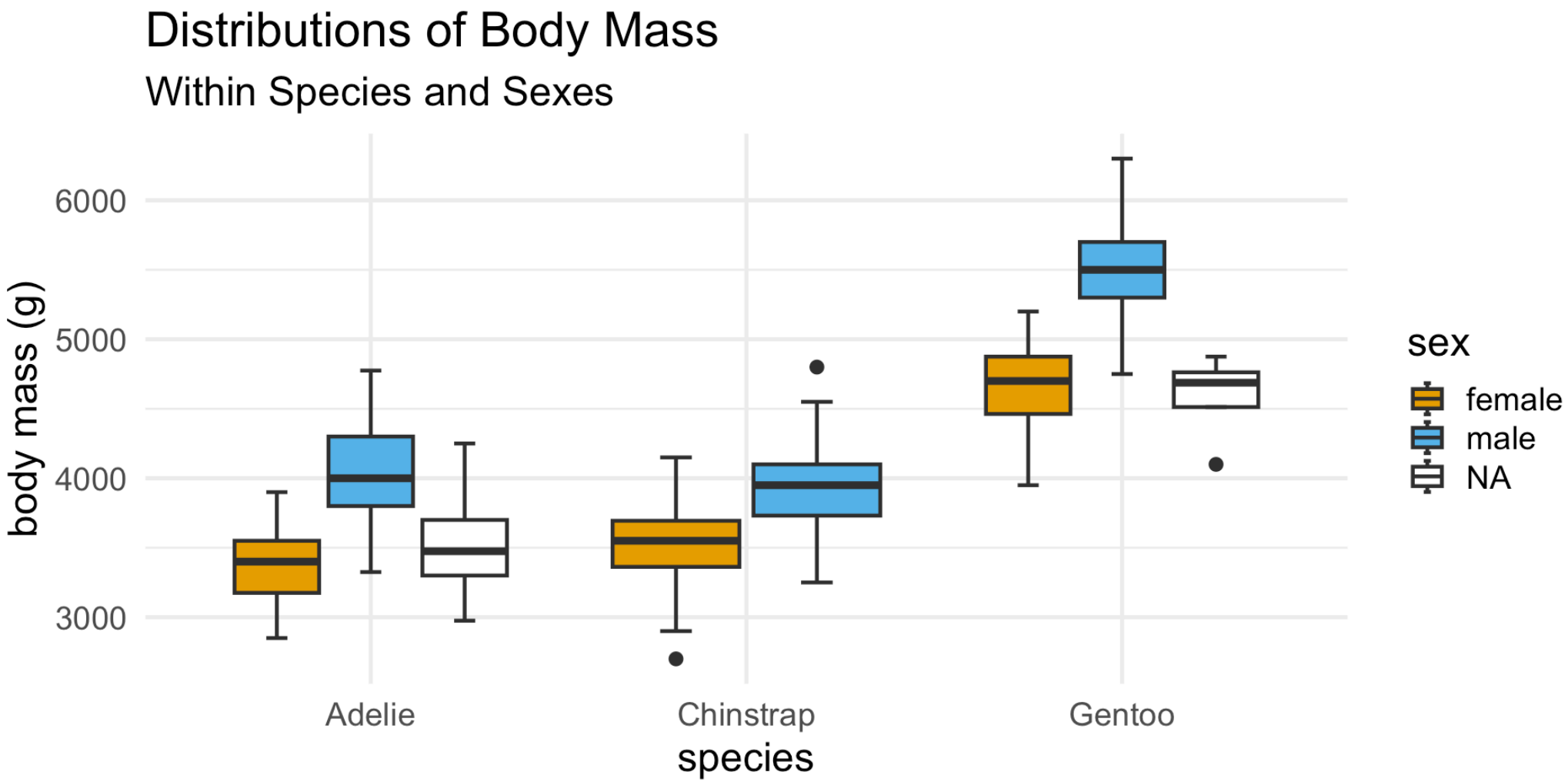
Second Idea: Use different shapes for different species:



- Key takeaway: we can encode information from additional variables by modifying certain attributes about the objects on our plots.

Penguins

Example



The Grammar of Graphics

Introduction

- Though we can make graphs “by hand” (with pen and paper), how can we tell a *computer* to make a graph?

  → To answer this question, we need to establish a framework with which we can decompose a plot into its constituent parts.

- Several such frameworks exist; one of the most popular is the **Grammar of Graphics**

  → First proposed by Leland Wilkinson in 1999, and then modified by Hadley Wickham in the 2000s

- We start with **data** (often in **tidy** format).

The Grammar of Graphics

Introduction

- Then, we need to specify **axes** / a **coordinate system**

  → What variable goes on the *x*-axis? What about the *y*-axis? Should we include a radial axis? Should we make a map?

- Finally, we need **geometric objects** (shortened to **geoms**)

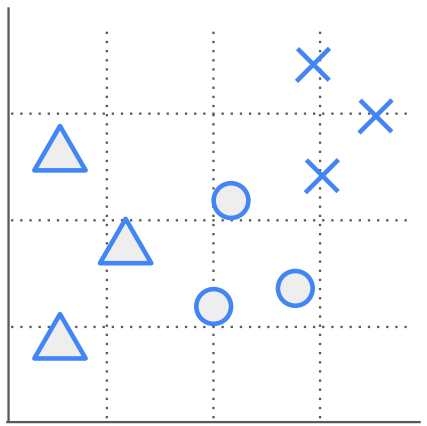
  → Do we need bars or points? Lines or sectors? Etc.

- **Aesthetics** are additional attributes of the geoms, to which variables can be mapped (e.g. coordinates of points, heights of bars, etc.)

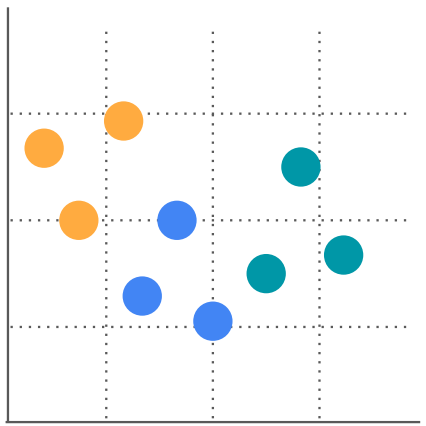
  → Be careful to distinguish the aesthetics from the aesthetic *mappings* - the latter is what maps the data to the former.

The Grammar of Graphics

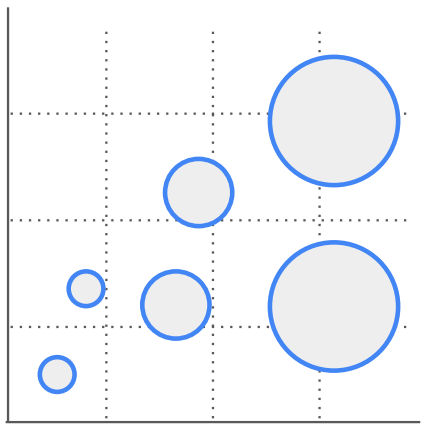
Some Common Aesthetics



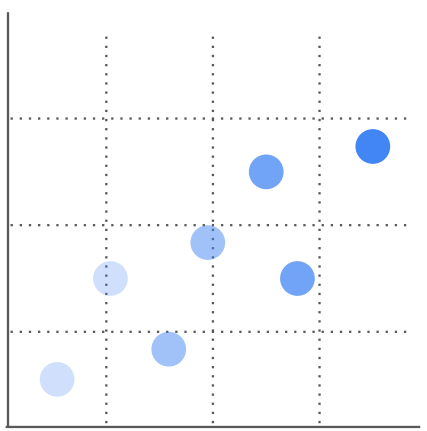
Shapes (points)



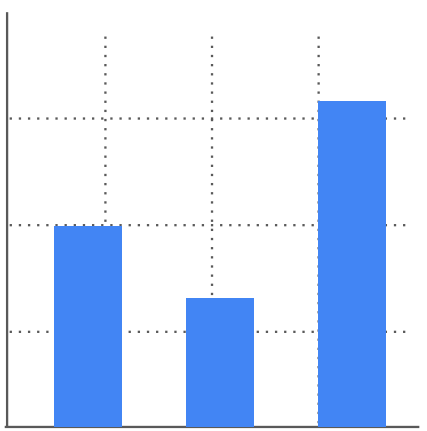
Color



Size



Opacity

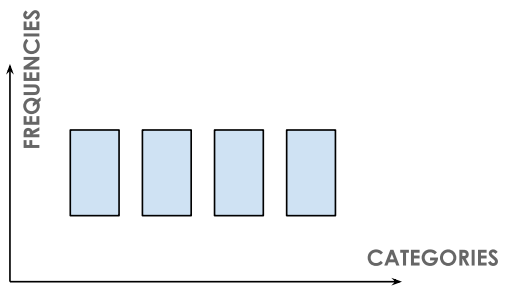


Height

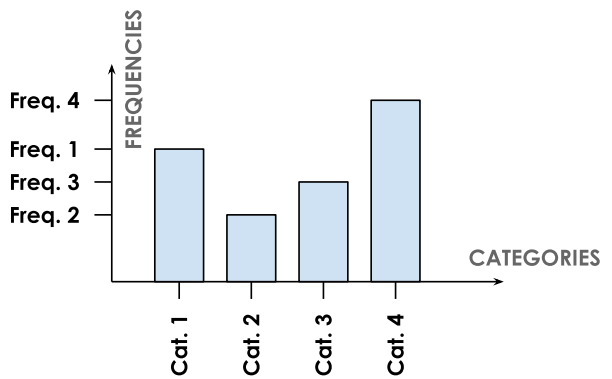
The Grammar of Graphics

Example

SPECIFY AXES



ADD GEOMS (rectangles)



SPECIFY AESTHETIC MAPPINGS

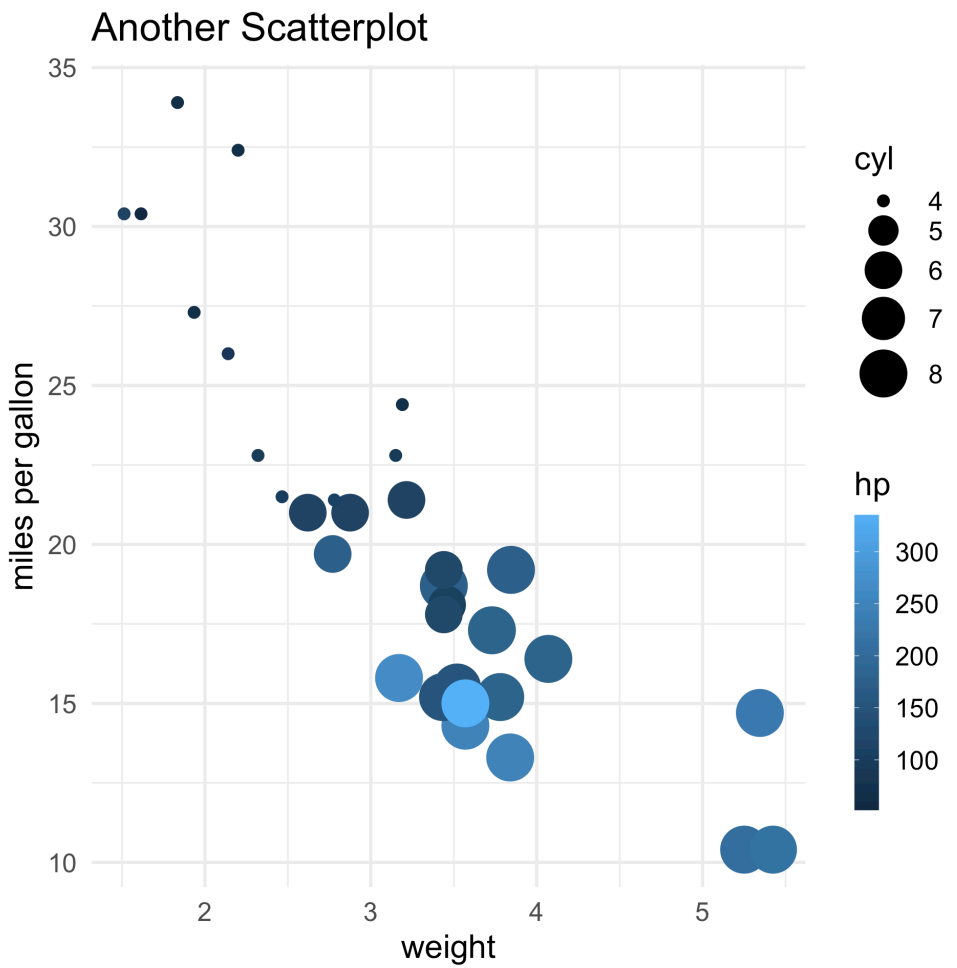
- x-coord ↦ categories
- height ↦ frequency

Aesthetics From Data

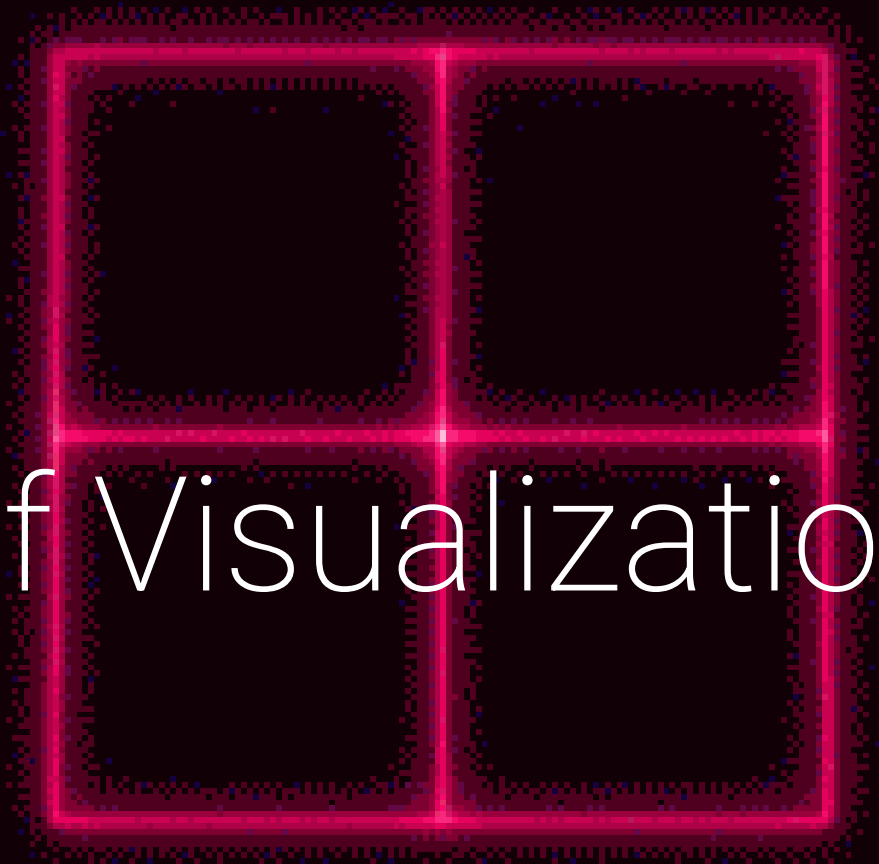
- Note that (at least for this workshop), we are treating *x*- and *y*-coordinates as aesthetics.

mtcars

Check Your Understanding



- How many variables are being compared?
- What aesthetic is each mapped to?
- What conclusions can we draw from the plot?



Theory of Visualization

Principles of Good Visualization

Setting Goals

- When setting out to make a plot, it's important to be intentional about our goals.
- There are two main types of plots: exploratory, and presentation-quality.

Exploratory	Presentation-Quality

- Summarize trends/patterns before performing more sophisticated statistical analyses
- Details not too important; quantity over quality

Principles of Good Visualizations

Tips and Tricks

Here are some tips I've found useful when crafting visualizations

1. **Keep things simple.** You can (and in many cases should) try to communicate as much information as is effective. But, don't take it to an extreme.
    - **3D-Styling is almost NEVER effective**. As neat and "cool" as 3D barplots might be, the 3D-styling elements often obfuscate the plot's true meaning
2. **Beware of Scales and Areas**. We;ll talk about this one more in a bit - spoiler alert, pie charts are a notoriously bad graphic!

Principles of Good Visualizations

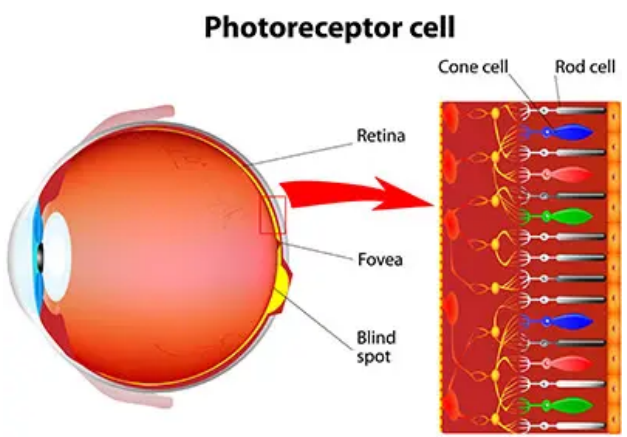
Tips and Tricks

3. **Label Axes, and Title your Plots.** This one should (hopefully) be self-explanatory, but make sure you are using descriptive (but not overly complex) labels for your axes, and make sure your plots are titled.
4. **Interpret your plots.** All too often I see “floating” plots - that is, figures that appear mysteriously and suddenly with no explanation whatsoever. No matter how self-explanatory you think your plot is, make sure you actively describe it and its conclusions somewhere in your report.

- There’s a bit more I’d like to say on the use of **color** as well.

CVD and Accessibility

- Especially when it comes to presentation-oriented graphics, accessibility is key.
- One thing to keep in mind that many readers may suffer from Color-Vision Deficiency (CVD; aka colorblindness), and may not be able to easily perceive differences in colors.
  - → **Deuteranomaly**: difficulty perceiving green
  - → **Protanomaly**: difficulty perceiving red
  - → **Tritanomaly**: difficulty perceiving blue



Trichromatic persons (i.e. people with no colorblindness) possess all three retinal cone cell types (and have cone cell types that function "as expected", and are therefore able to process and perceive red, green, and blue light

Image Source: https://www.aao.org/eye-health/anatomy/cones

CVD and Accessibility

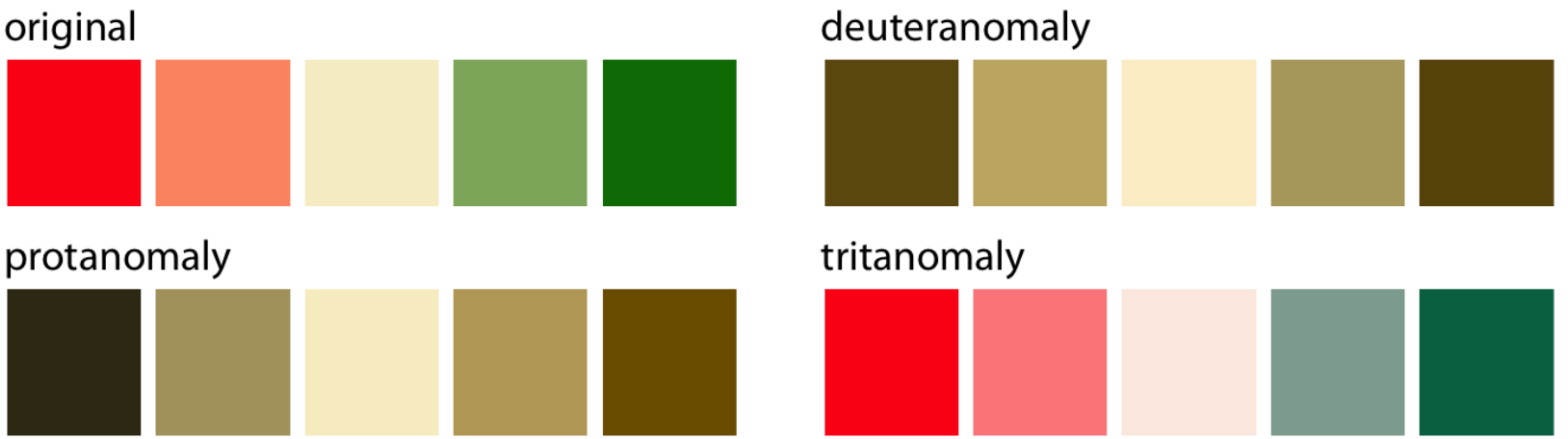


Figure 19.7 from *Fundamentals of Data Visualization*: A red–green contrast becomes indistinguishable under red–green cvd (deuteranomaly or protanomaly).

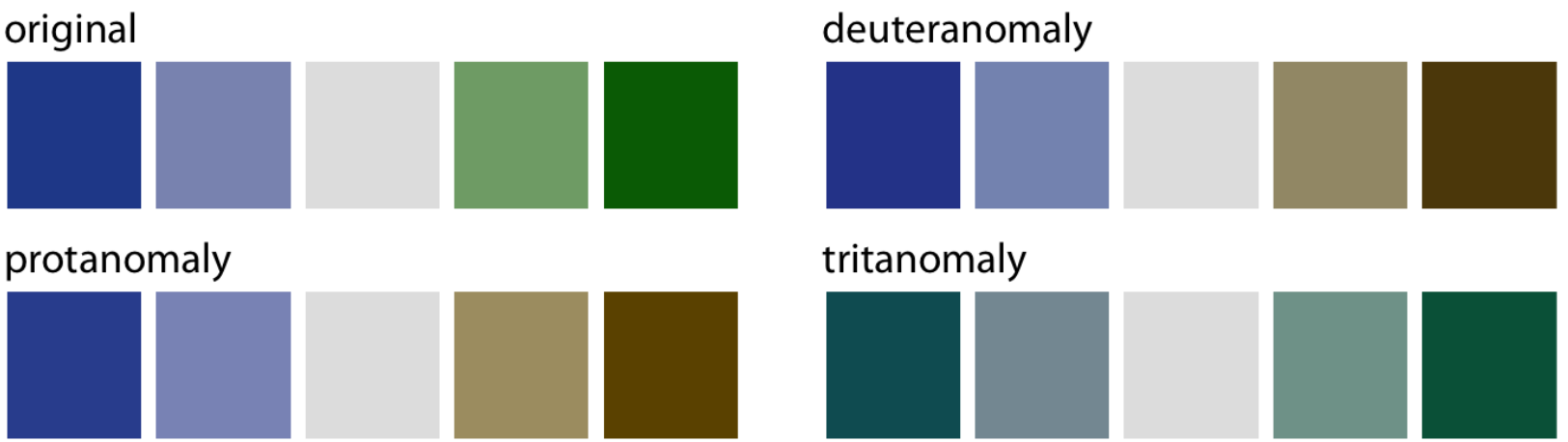


Figure 19.8 from *Fundamentals of Data Visualization*: A blue–green contrast becomes indistinguishable under blue–yellow cvd (tritanomaly).

CVD and Accessibility

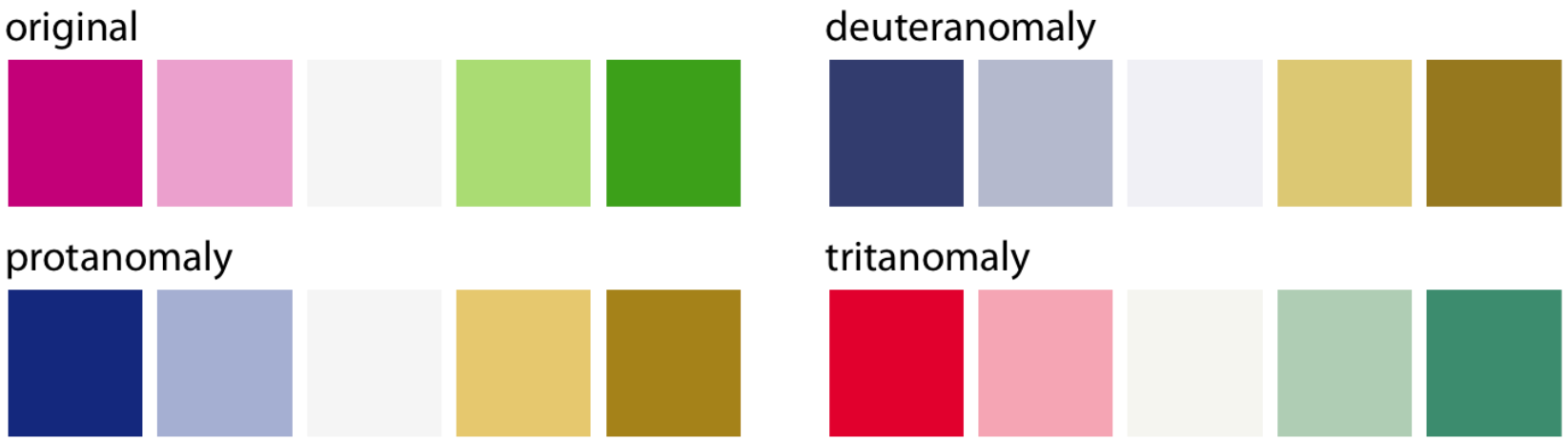


Figure 19.9 from *Fundamentals of Data Visualization*: The ColorBrewer PiYG (pink to yellow-green) scale from Figure 4.5 looks like a red–green contrast to people with regular color vision but works for all forms of color-vision deficiency. It works because the reddish color is actually pink (a mix of red and blue) while the greenish color also contains yellow. The difference in the blue component between the two colors can be picked up even by deutans or protans, and the difference in the red component can be picked up by tritans.

CVD and Accessibility

The Okabe-Ito Palette

Figure 19.10 from *Fundamentals of Data Visualization*: Qualitative color palette for all color-vision deficiencies (Okabe and Ito 2008). The alphanumeric codes represent the colors in RGB space, encoded as hexadecimals. In many plot libraries and image-manipulation programs, you can just enter these codes directly. If your software does not take hexadecimals directly, you can also use the values in Table 19.1.

```
1 palette.colors(palette = "Okabe-Ito")
```

```
[1] "#000000" "#E69F00" "#56B4E9" "#009E73" "#F0E442" "#0072B2"
"#D55E00"
[8] "#CC79A7" "#999999"
```

- Another resource: https://www.color-blindness.com/coblis-color-blindness-simulator/

Color Scales

Three Main Types

- It is also important to make sure you are using a **color scale** that is appropriate for your visualization

  → Loosely speaking, you can think of a "color scale" as a palette of colors that will appear on your plot.

- There are three main types of color scales:

  → **Qualitative**: colors are distinct, with no natural order. Good for use with categorical variables.

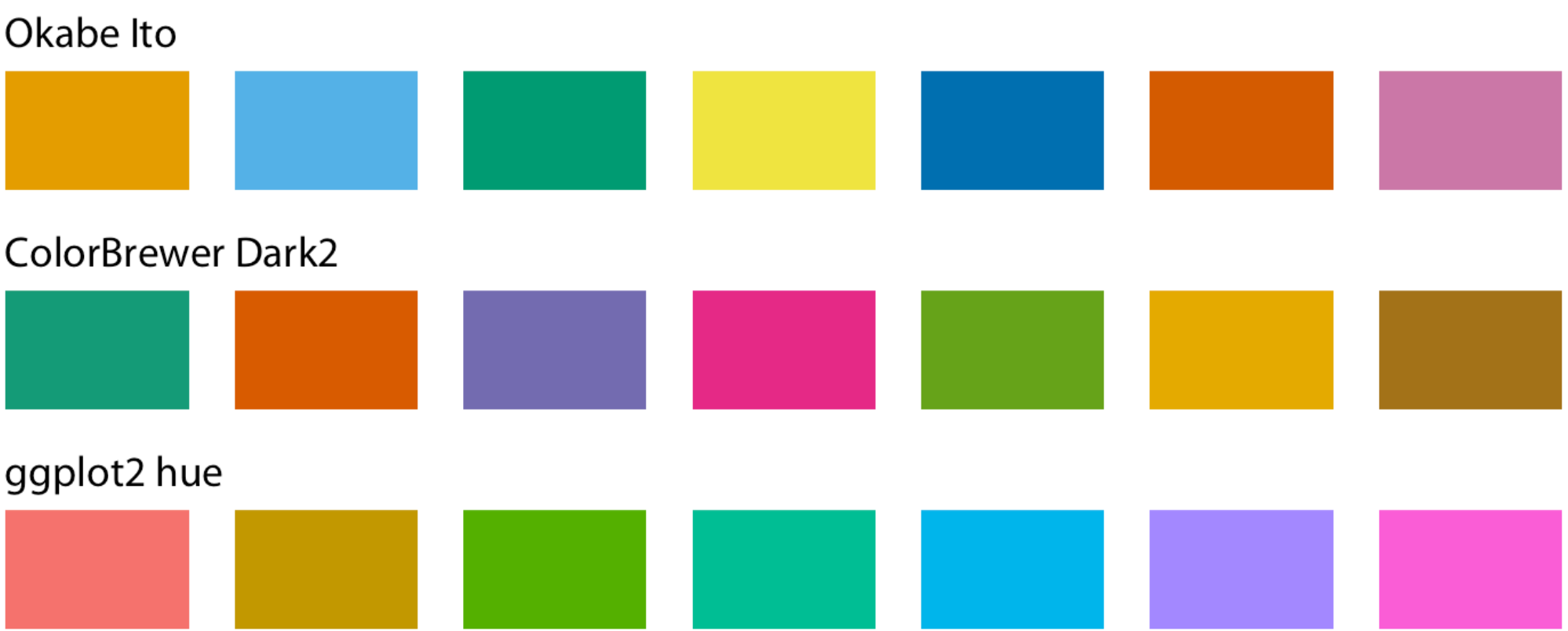
  → **Sequential**: colors range from light to dark, and are used to convey a *direction*. Similar to what we colloquially call "gradients"

  → **Diverging:**: two sequential scales stitched together at a neutral midpoint.

Color Scales

Three Main Types

Qualitative | Sequential | Diverging

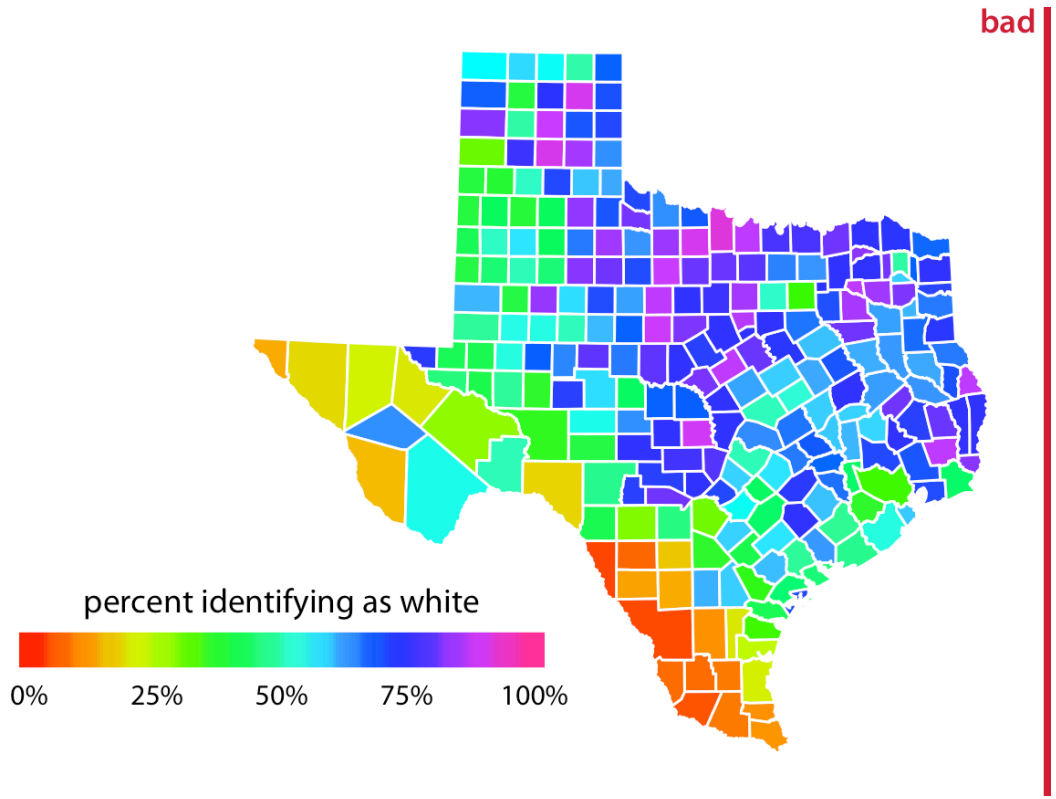


Source: *Fundamentals of Data Visualization, by Claus Wilke*

Color Scales

Example

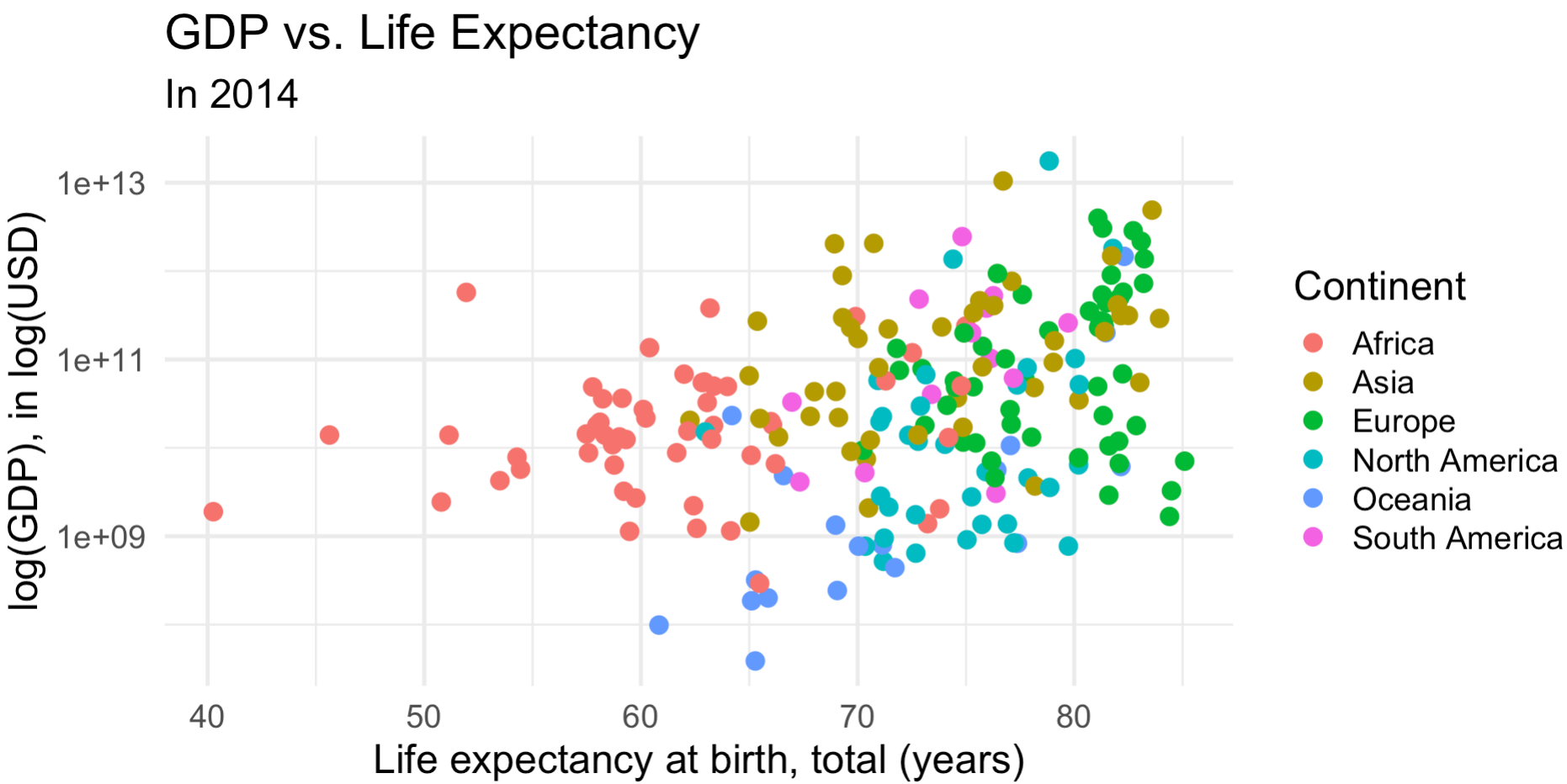
Misuse | Improvement



Source: *Fundamentals of Data Visualization, by Claus Wilke*

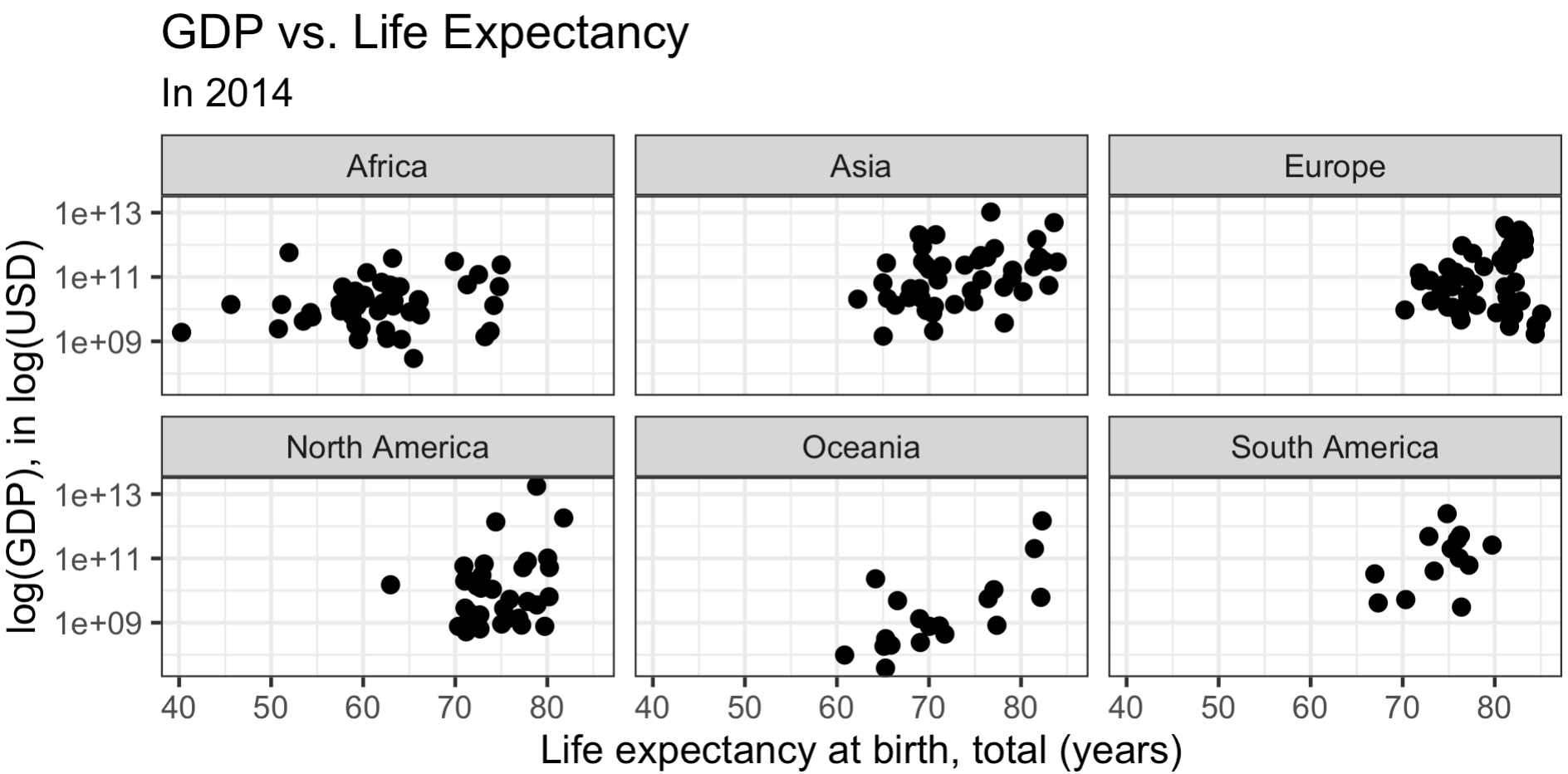
Facetting

- Color may not always be the most effective way to convey information.



Facetting

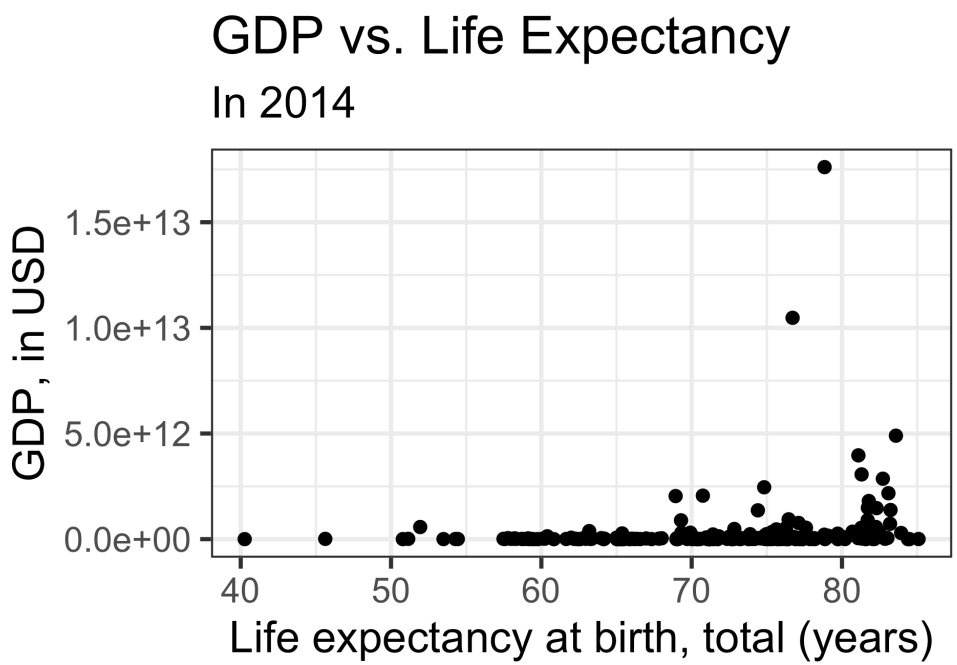
- One potential alternative is **facetting**



Transformations

- Also note how **transformations** may be useful, especially when one or more of your variables has comparatively high spread.

Raw:



Log-Transformed: