Lecture 2. Visual Vocabulary & Effective Visualizations

PUBH 6199: Visualizing Data with R, Summer 2025

Xindi (Cindy) Hu, ScD

2025-05-27

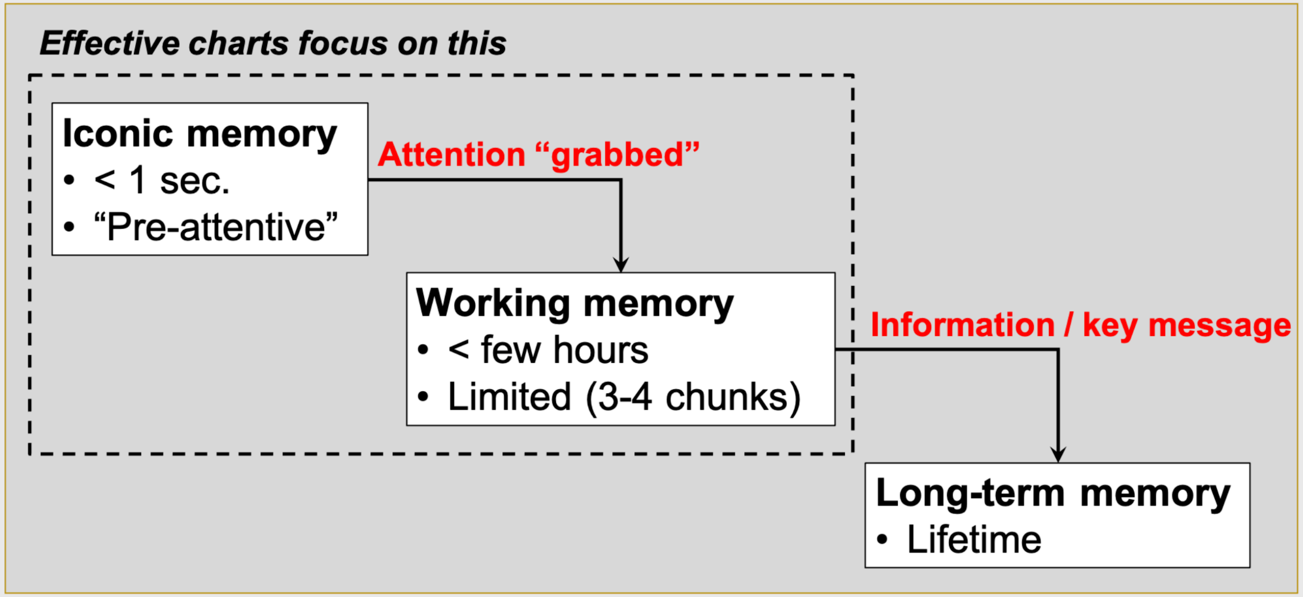
Outline for today

- **How human see data**
- Data-Ink Maximization and Graphical Redesign
- Design considerations for different types of intended audience

Good data visualization is optimized for our visual-memory system

- Helps us **understand trends and patterns**
- Makes data **more accessible** to different audiences
- Useful in **decision-making** and **communication**

A (very) simplified model of the visual-memory system



The power of pre-attentive processing

Count all the 5s in the following image

821134907856412043612

304589640981709812734

123450986124790812734

029860192837401489363

123479827961203459816

234009816256908127634

123459087162342015237

123894789237498230192

The power of pre-attentive processing

Count all the 5s in the following image

8211349078**5**6412043612

304**5**89640981709812734

1234**5**0986124790812734

029860192837401489363

1234798279612034**5**9816

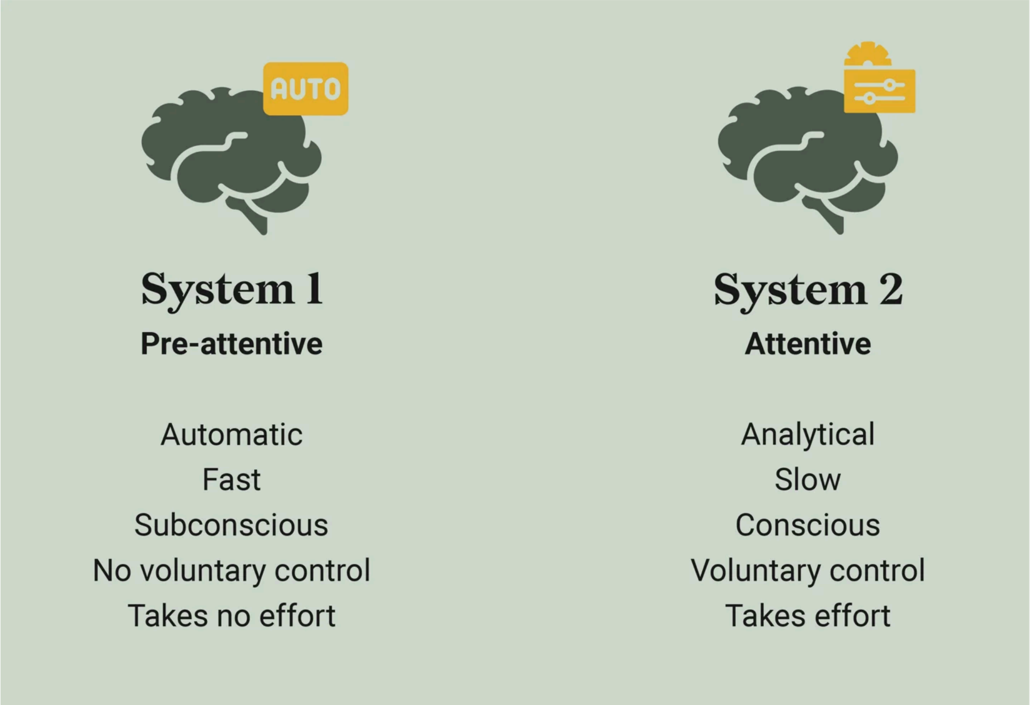
2340098162**5**6908127634

1234**5**908716234201**5**237

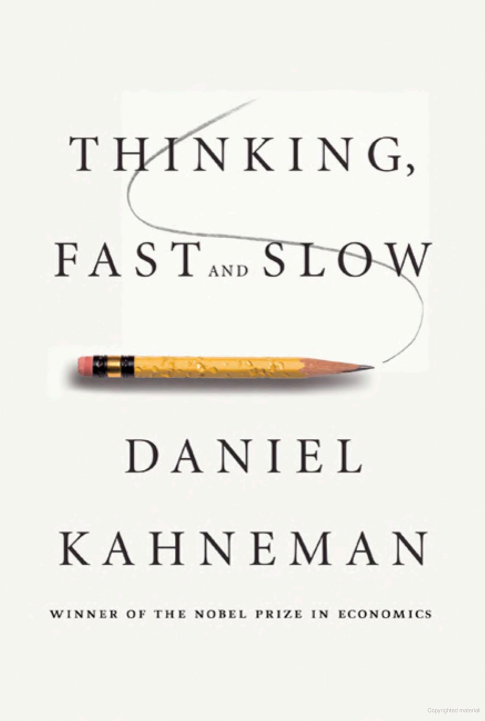
123894789237498230192

What is pre-attentive processing?

- Rapid, automatic processing of visual information before conscious attention kicks in.
- Happens within <250 milliseconds.
- Helps identify key patterns without effort.

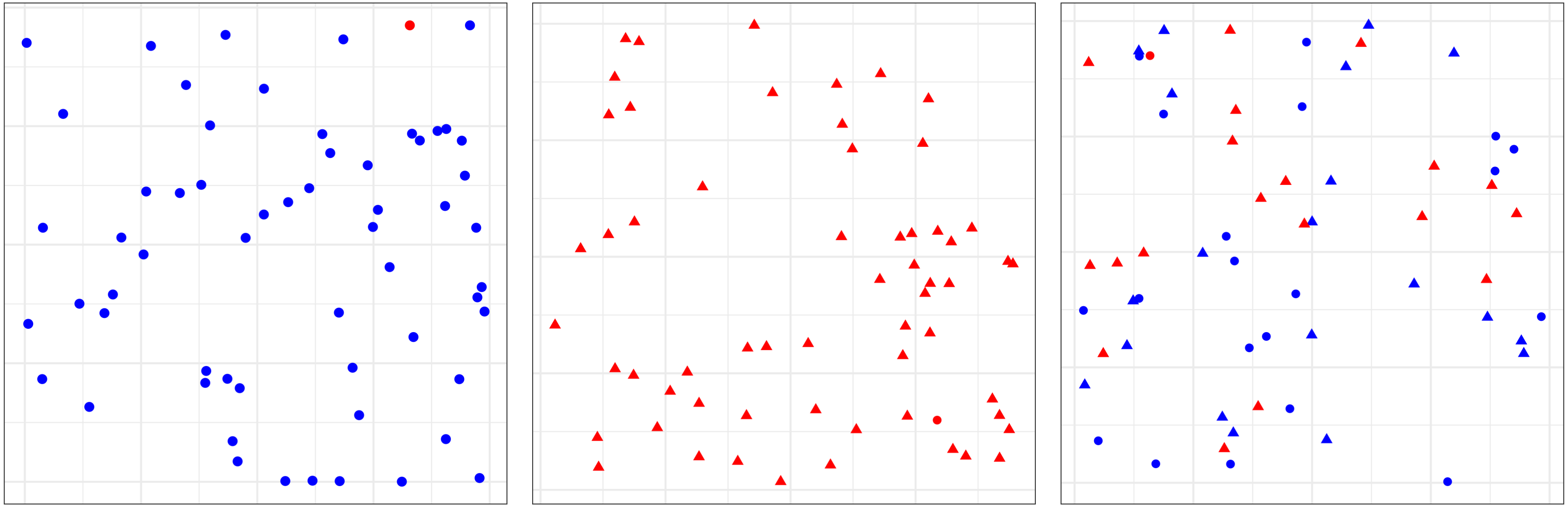


System 1	System 2
Pre-attentive	Attentive
Automatic	Analytical
Fast	Slow
Subconscious	Conscious
No voluntary control	Voluntary control
Takes no effort	Takes effort

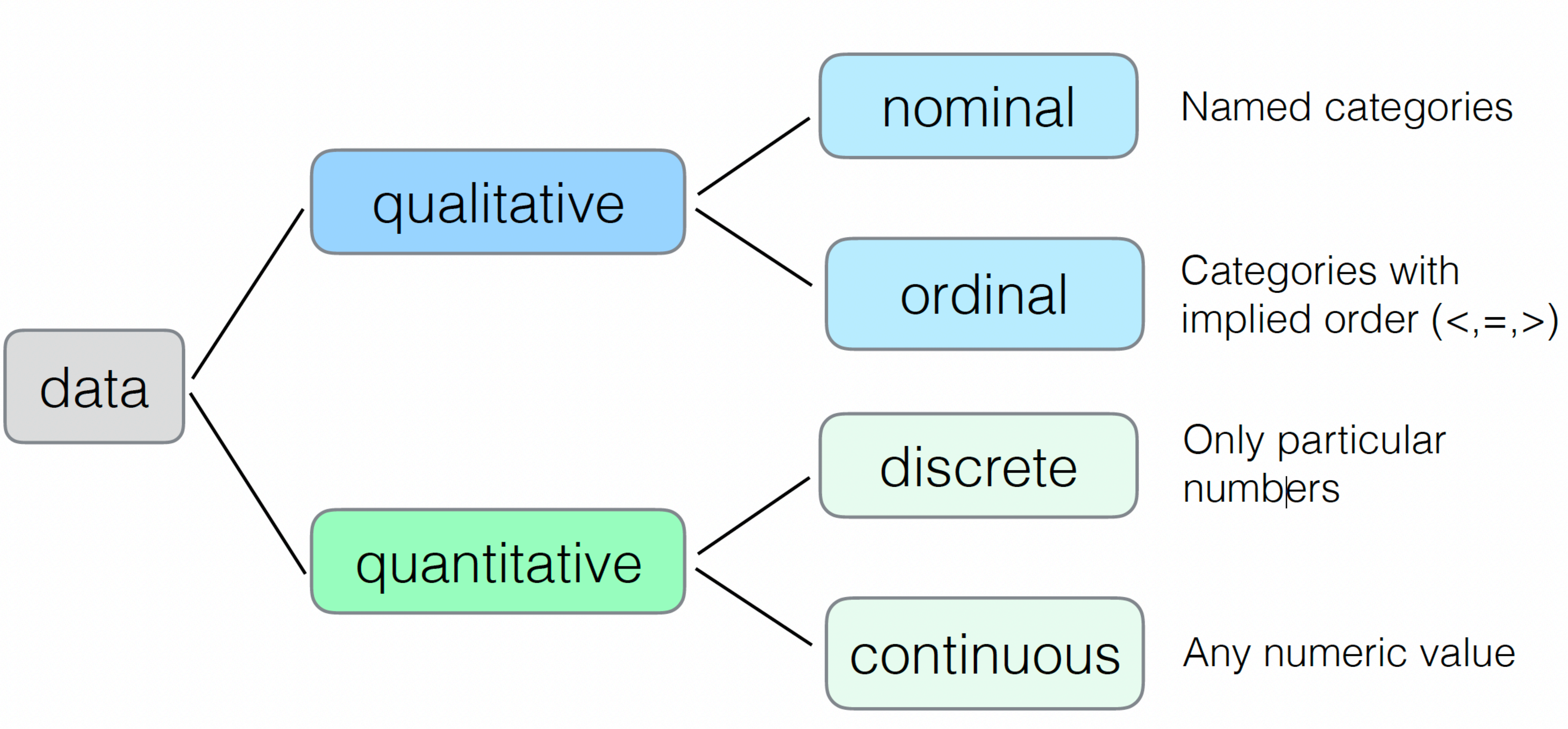


Not all pre-attentive features are created equal

Raise your hand when you see the red dot?



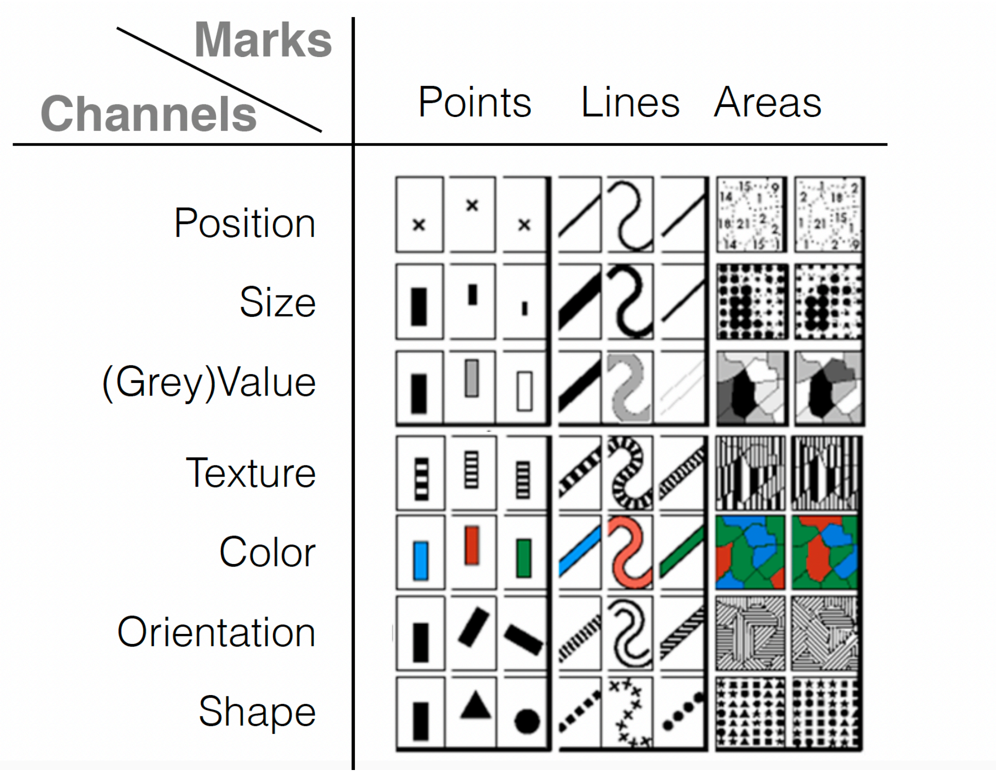
Classify data types

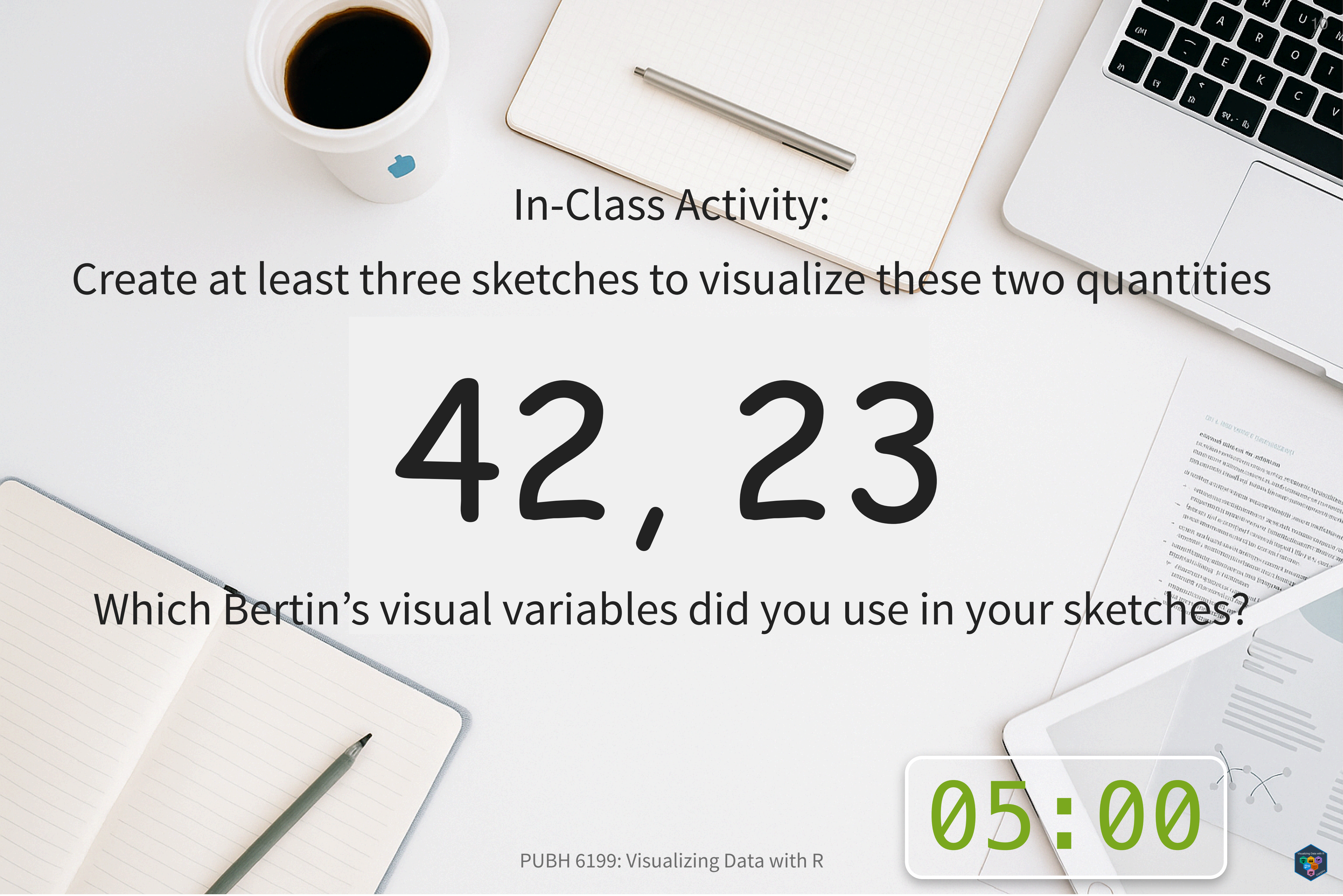


Introducing visual variable

> “A **visual variable**, in data visualization, is an aspect of a graphical object that can visually differentiate it from other objects, and can be controlled during the design process.”

- Jacques Bertin, 1967, *Sémiologie Graphique*





In-Class Activity:

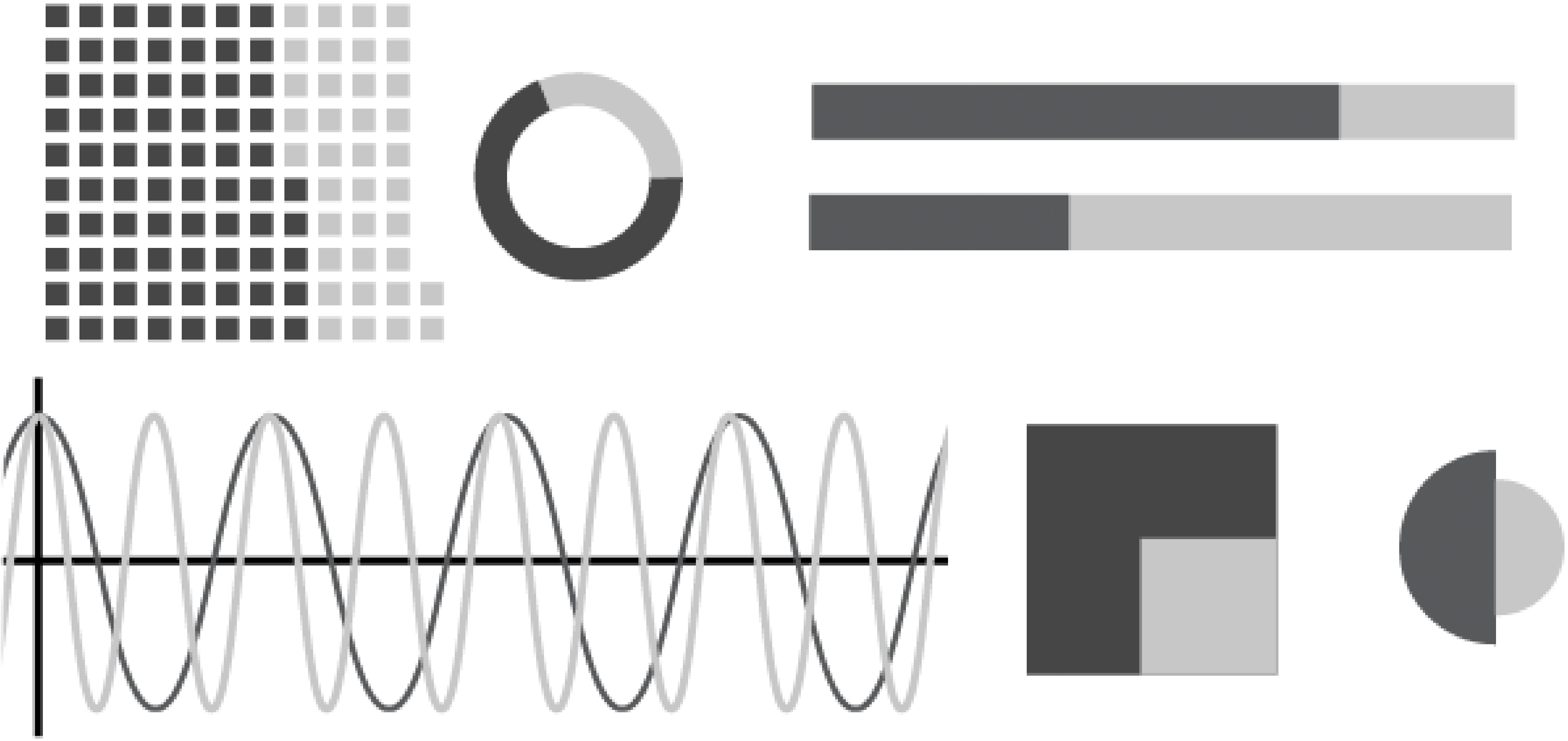
Create at least three sketches to visualize these two quantities

42, 23

Which Bertin’s visual variables did you use in your sketches?

05:00

45 ways to visualizae two quantities



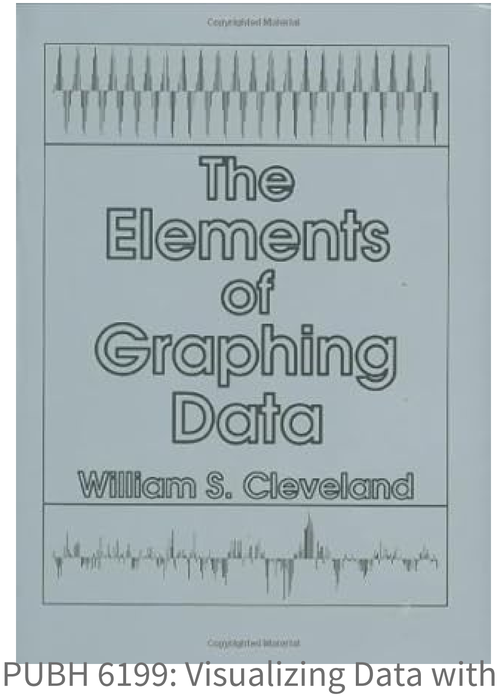
https://rockcontent.com/blog/45-ways-to-communicate-two-quantities/

Cleveland’s three visual operations of pattern perception

🎯 **Detection**: *Recognizing that a geometric object encodes a physical value.*

🧩 **Assembly**: *Grouping detected graphical elements into patterns.*

📏 **Estimation**: *Visually assessing the relative magnitude of two or more values.*



Copyrighted Material

The Elements of Graphing Data

William S. Cleveland

Copyrighted Material

Graphical Perception and Graphical Methods for Analyzing Scientific Data

William S. Cleveland and Robert McGill

Summary. Graphical perception is the visual decoding of the quantitative and qualitative information encoded on graphs. Recent investigations have uncovered basic principles of human graphical perception that have important implications for the display of data. The computer graphics revolution has stimulated the invention of many graphical methods for analyzing and presenting scientific data, such as box plots, two-tiered error bars, scatterplot smoothing, dot charts, and graphing on a log base 2 scale.

Graphs provide powerful tools both for analyzing scientific data and for communicating quantitative information. The computer graphics revolution, which began in the 1960's and has intensified during the past several years, stimulated the invention of graphical methods: types of graphs and types of quantitative information to be shown on graphs (*1–4*). One purpose of this article is to describe and illustrate several of these new methods.

What has been missing, until recently, in this period of rapid graphical invention and deployment is the study of graphs and the human visual system. When a graph is constructed, quantitative and categorical information is encoded, chiefly through position, shape, size, symbols, and color. When a person looks at a graph, the information is visually decoded by the person's visual system. A graphical method is successful only if the decoding is effective. No matter how clever and how technologically impressive the encoding, it fails if the decoding process fails. Informed decisions about how to encode data can be achieved only through an understanding of this visual decoding process, which we call graphical perception (*5*).

Our second purpose is to convey some recent theoretical and experimental investigations of graphical perception. We identify certain elementary graphical-perception tasks that are performed in the visual decoding of quantitative information from graphs; theory and experimental data are then used to order the tasks on the basis of accuracy. The ordering has an important application: data should be encoded so that the visual decoding involves tasks as high in the ordering as possible, that is, tasks performed with greater accuracy. This is illustrated by several examples in which some much-used graphical forms are presented, set aside, and replaced by new methods.

Elementary Tasks for the Graphical Perception of Quantitative Information

The first step is to identify elementary graphical-perception tasks that are used to visually extract quantitative information from a graph. (By ''quantitative information'' we mean numerical values of a variable, such as frequency of radiation and gross national product, that are not highly discrete; this excludes categorical information, such as type of metal and nationality, which is also shown on many graphs.) Ten tasks with which we have worked, in our theoretical investigations and in our experiments, are the following: angle, area, color hue, color saturation, density (amount of black), length (distance), position along a common scale, positions on identical but nonaligned scales, slope, and volume (Fig. 1).

Visual decoding as we define it for elementary graphical-perception tasks is what Julesz calls preattentive vision (*6*): the instantaneous perception of the visual field that comes without apparent mental effort. We also perform cognitive tasks such as reading scale information, but much of the power of graphs—and what distinguishes them from tables—comes from the ability of our preattentive visual system to detect geometric patterns and assess magnitudes. We have examined preattentive processes rather than cognition.

We have studied the elementary graphical-perception tasks theoretically, borrowing ideas from the more general field of visual perception (*7, 8*), and experimentally by having subjects judge graphical elements (*1, 5*). The next two sections illustrate the methodology with a few examples.

Study of Graphical Perception: Theory

Figure 2 provides an illustration of theoretical reasoning that borrows some ideas from the field of computational vision (*8*). Suppose that the goal is to judge the ratio, r, of the slope of line segment BC to the slope of line segment AB in each of the three panels. Our visual system tells us that r is greater than 1 in each panel, which is correct. Our visual system also tells us that r is closer to 1 in the two rectangular panels than in the square panel; that is, the slope of BC appears closer to the slope of AB in the two rectangular panels than in the square panel. This, however, is incorrect; r is the same in all three panels.

The reason for the distortion in judging Fig. 2 is that our visual system is geared to judging angle rather than slope. In their work on computational theories of vision in artificial intelligence, Marr (*8*) and Stevens (*9*) have investigated how people judge the slant and tilt (*10*) of the surfaces of three-dimensional objects. They argue that we judge slant and tilt as angles and not, for example, as their tangents, which are the slopes. An angle contamination of slope judgments explains the distortion in judgments of Fig. 2. Let the angle of a line segment be the angle between it and a horizontal ray extending to the right (θ in Fig. 3). The angles of the line segments in the square panel of Fig. 2 are not as similar in magnitude as the angles in either of the rectangular panels; this makes the slopes in the rectangular panels seem closer in value.

Again, let θ be the angle of a line segment. Suppose a second line segment has an angle $\theta + \Delta\theta$ where $\Delta\theta$ is small but just large enough that a difference in the orientations of the line segments can

The authors are statistical scientists at AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974.

828

SCIENCE, VOL. 229

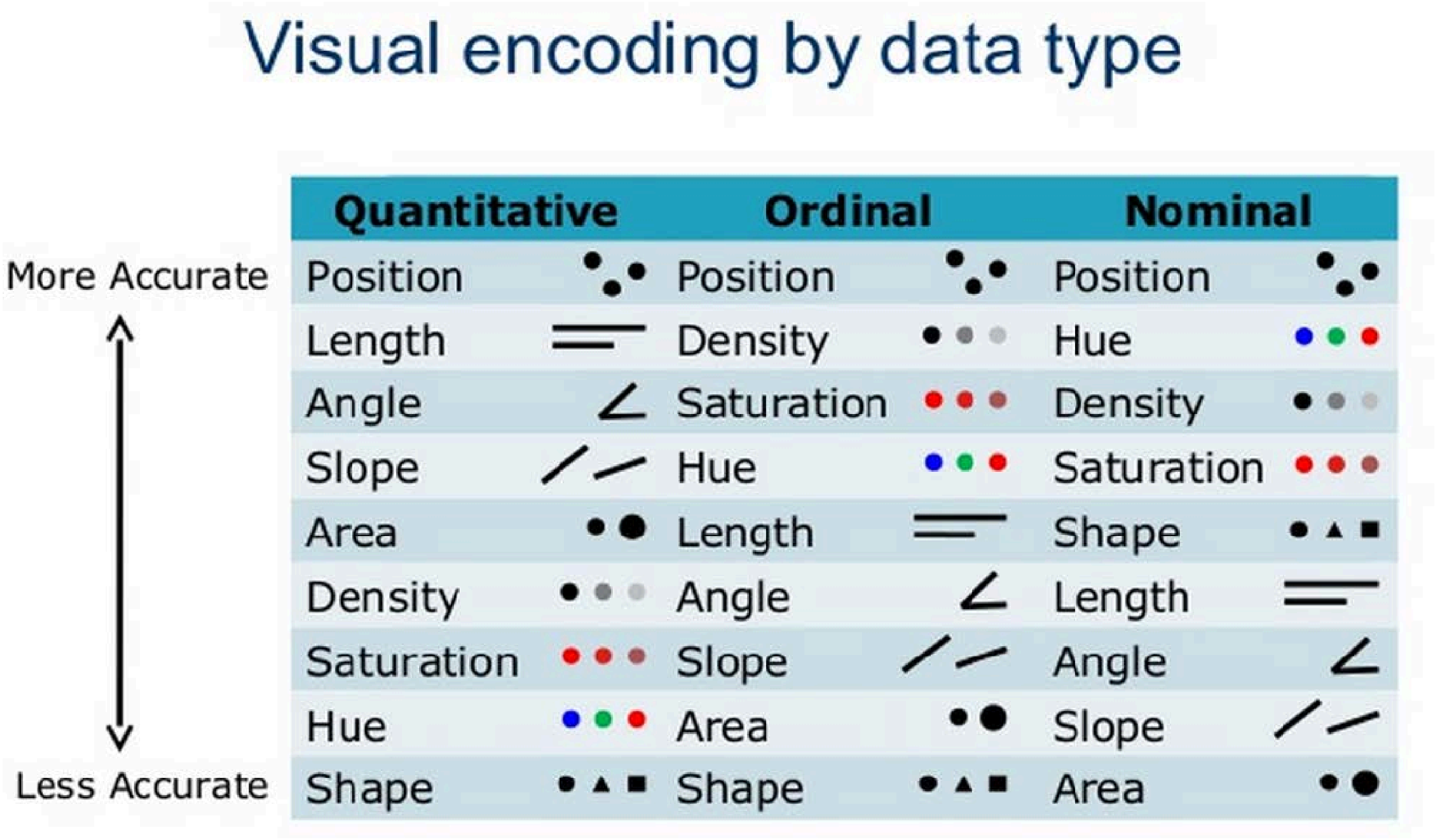
Starting with estimation because it is the hardest

Three levels of estimation

Level	Example
1. Discrimination	X = Y X != Y
2. Ranking	X < Y X > Y
3. Ratioing	X / Y = ?

📏 **We want to get as far down this list as possible with efficiency and accuracy**

What visual cues are most effective for which type of data?



Visual encoding by data type

	Quantitative	Ordinal	Nominal
More Accurate	Position	Position	Position
↑	Length	Density	Hue
	Angle	Saturation	Density
	Slope	Hue	Saturation
	Area	Length	Shape
	Density	Angle	Length
	Saturation	Slope	Angle
↓	Hue	Area	Slope
Less Accurate	Shape	Shape	Area

Source: Yau, N. (2013). Data Points: Visualization That Means Something. Wiley.

Introducing the coffee ratings dataset

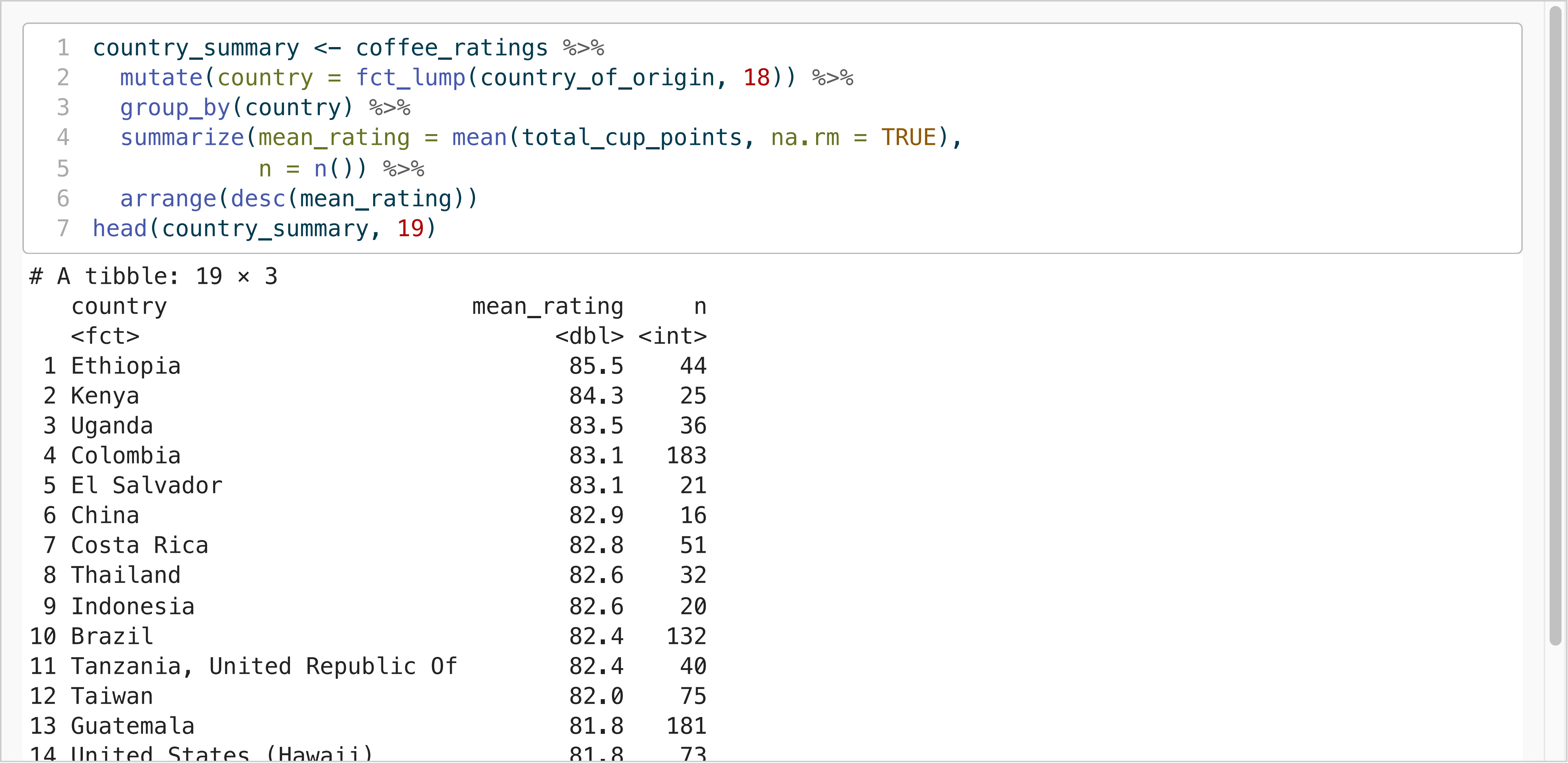
- These data contain reviews of 1312 arabica and 28 robusta coffee beans from the **Coffee Quality Institute**’s trained reviewers. (Link to dataset)
- It contains detailed information on coffee samples from different countries, focusing on nine attributes like **aroma, flavor, aftertaste, acidity, body, balance, uniformity, cup cleanliness, sweetness**.
- **Total cup points** measures the overall coffee quality.

```r
library(tidyverse)
library(kableExtra)
coffee_ratings <- readr::read_csv("data/coffee_ratings.csv")
glimpse(coffee_ratings)
```

```
Rows: 1,337
Columns: 43
$ total_cup_points      <dbl> 90.58, 89.92, 89.75, 89.00, 88.83, 88.83, 88.75,…
$ species               <chr> "Arabica", "Arabica", "Arabica", "Arabica", "Ara…
$ owner                 <chr> "metad plc", "metad plc", "grounds for health ad…
$ country_of_origin     <chr> "Ethiopia", "Ethiopia", "Guatemala", "Ethiopia",…
$ farm_name             <chr> "metad plc", "metad plc", "san marcos barrancas …
$ lot_number            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
$ mill                  <chr> "metad plc", "metad plc", NA, "wolensu", "metad …
$ ico_number            <chr> "2014/2015", "2014/2015", NA, NA, "2014/2015", N…
$ company               <chr> "metad agricultural developmet plc", "metad agri…
```

Calculate country-level summaries

For each country in the 18 most frequent levels, calculate the average total cup points and the number of coffee bean varieties, lump the other countries into the `Other` category.



```r
country_summary <- coffee_ratings %>%
  mutate(country = fct_lump(country_of_origin, 18)) %>%
  group_by(country) %>%
  summarize(mean_rating = mean(total_cup_points, na.rm = TRUE),
            n = n()) %>%
  arrange(desc(mean_rating))
head(country_summary, 19)
```

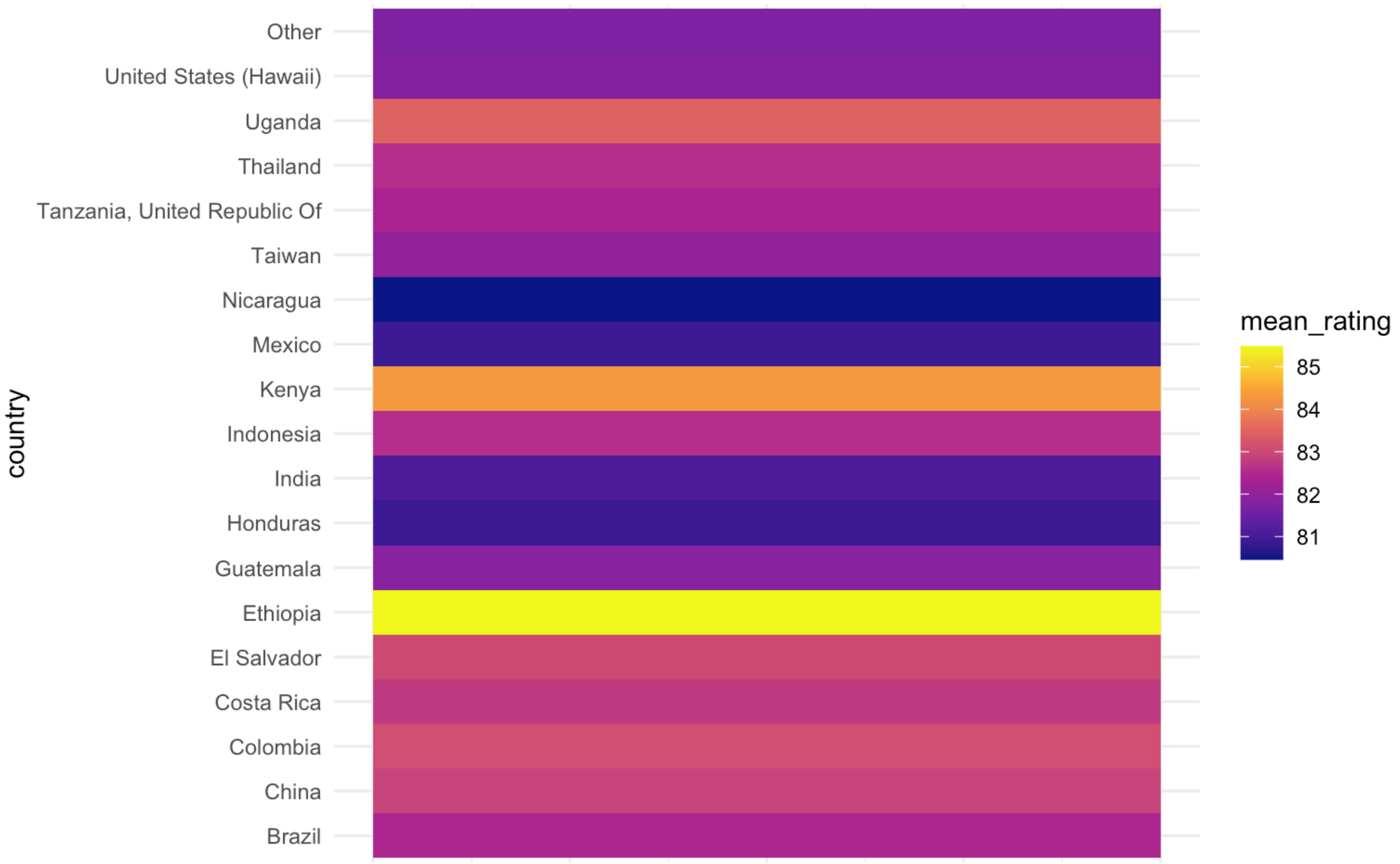
```
# A tibble: 19 × 3
   country                      mean_rating     n
   <fct>                              <dbl> <int>
 1 Ethiopia                            85.5    44
 2 Kenya                               84.3    25
 3 Uganda                              83.5    36
 4 Colombia                            83.1   183
 5 El Salvador                         83.1    21
 6 China                               82.9    16
 7 Costa Rica                          82.8    51
 8 Thailand                            82.6    32
 9 Indonesia                           82.6    20
10 Brazil                              82.4   132
11 Tanzania, United Republic Of        82.4    40
12 Taiwan                              82.0    75
13 Guatemala                           81.8   181
14 United States (Hawaii)              81.8    73
```

Let’s start from the bottom of the list

1. Position on a common scale
2. Position on non-aligned scales
3. Length
4. Angle
5. Area
6. Volume <> Density <> Color saturation
7. Color hue

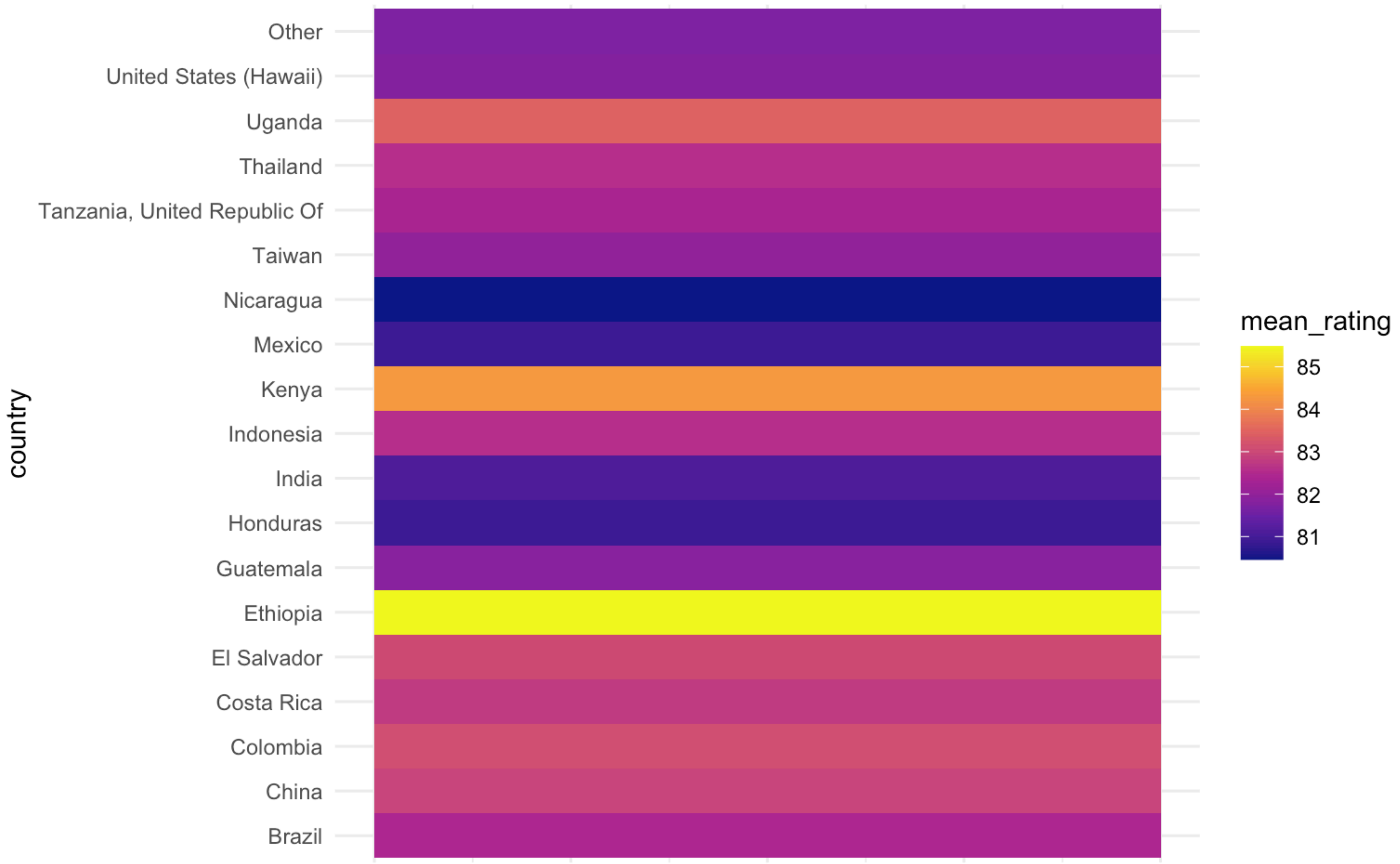
Use color hue to visualize average ratings

Easy: which has higher ratings, Kenya or Indonesia?

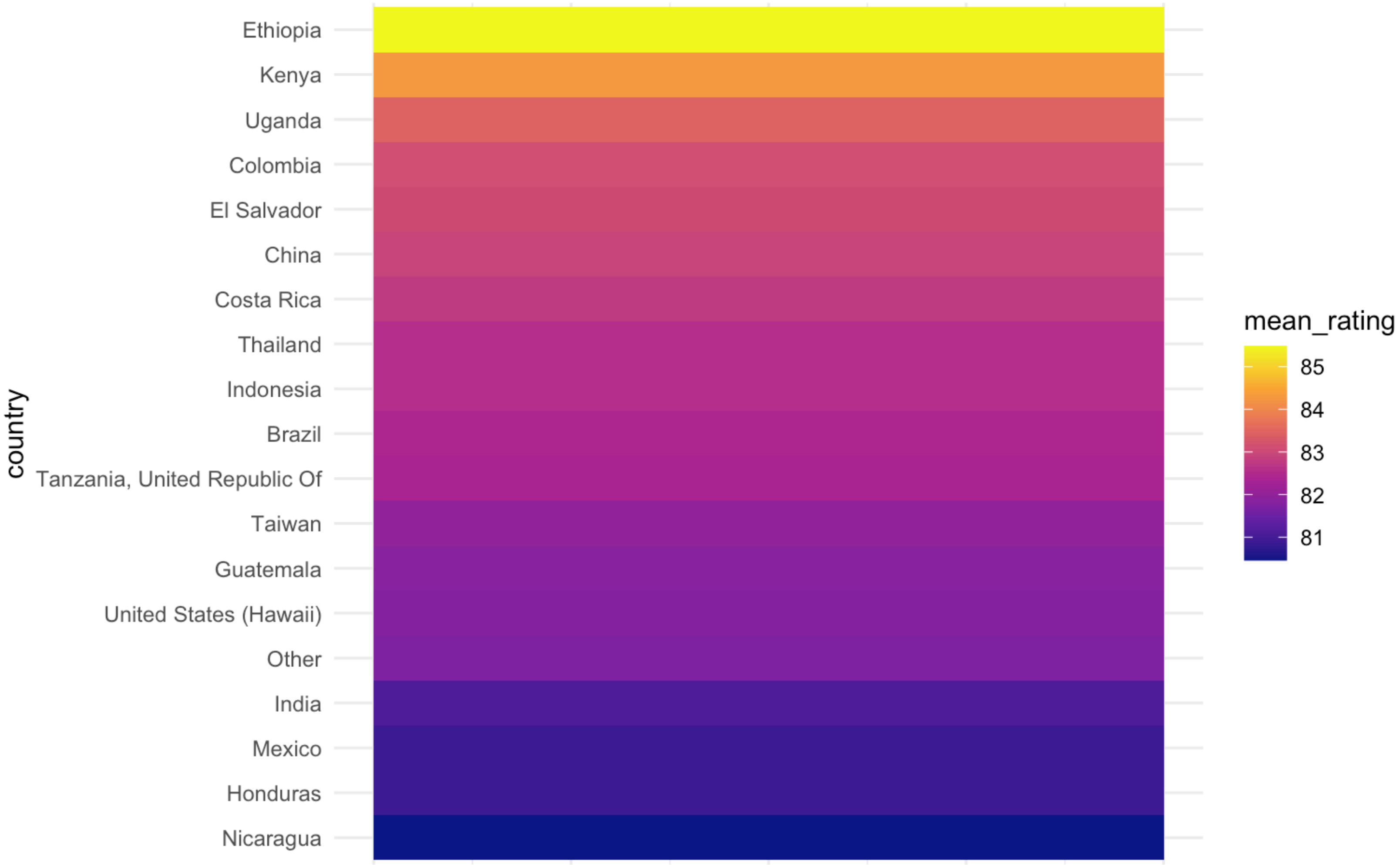


Use color hue to visualize average ratings

Hard: which has higher ratings, Indonesia or Costa Rica?



What about now?

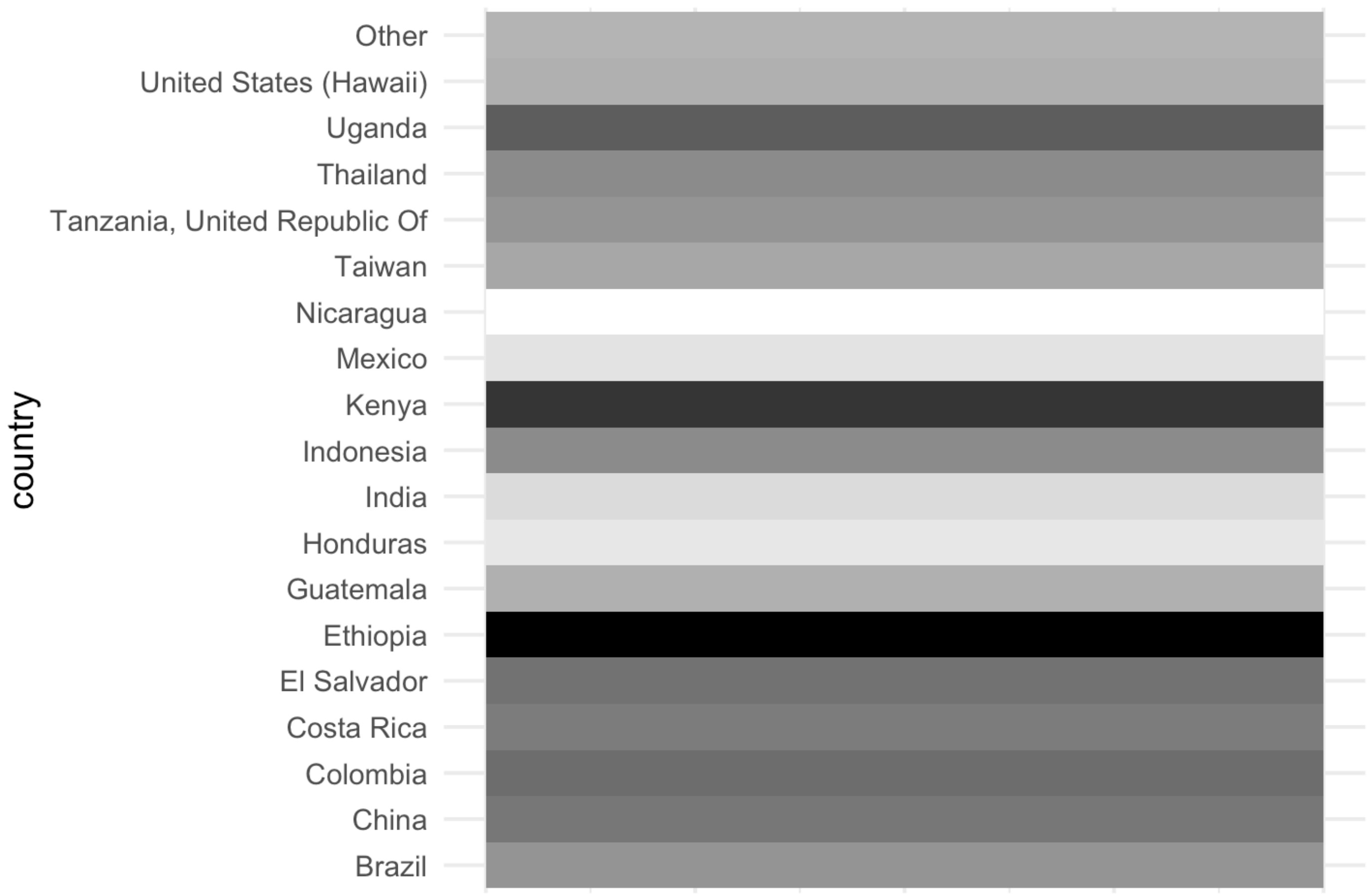


Observation: alphabetical ordering of the categorical variable is almost never useful, re-rank as needed.

Move up one level to color saturation

1. Position on a common scale
2. Position on non-aligned scales
3. Length
4. Angle
5. Area
6. Volume <> Density <> Color saturation
7. Color hue

Use color saturation to visualize average ratings

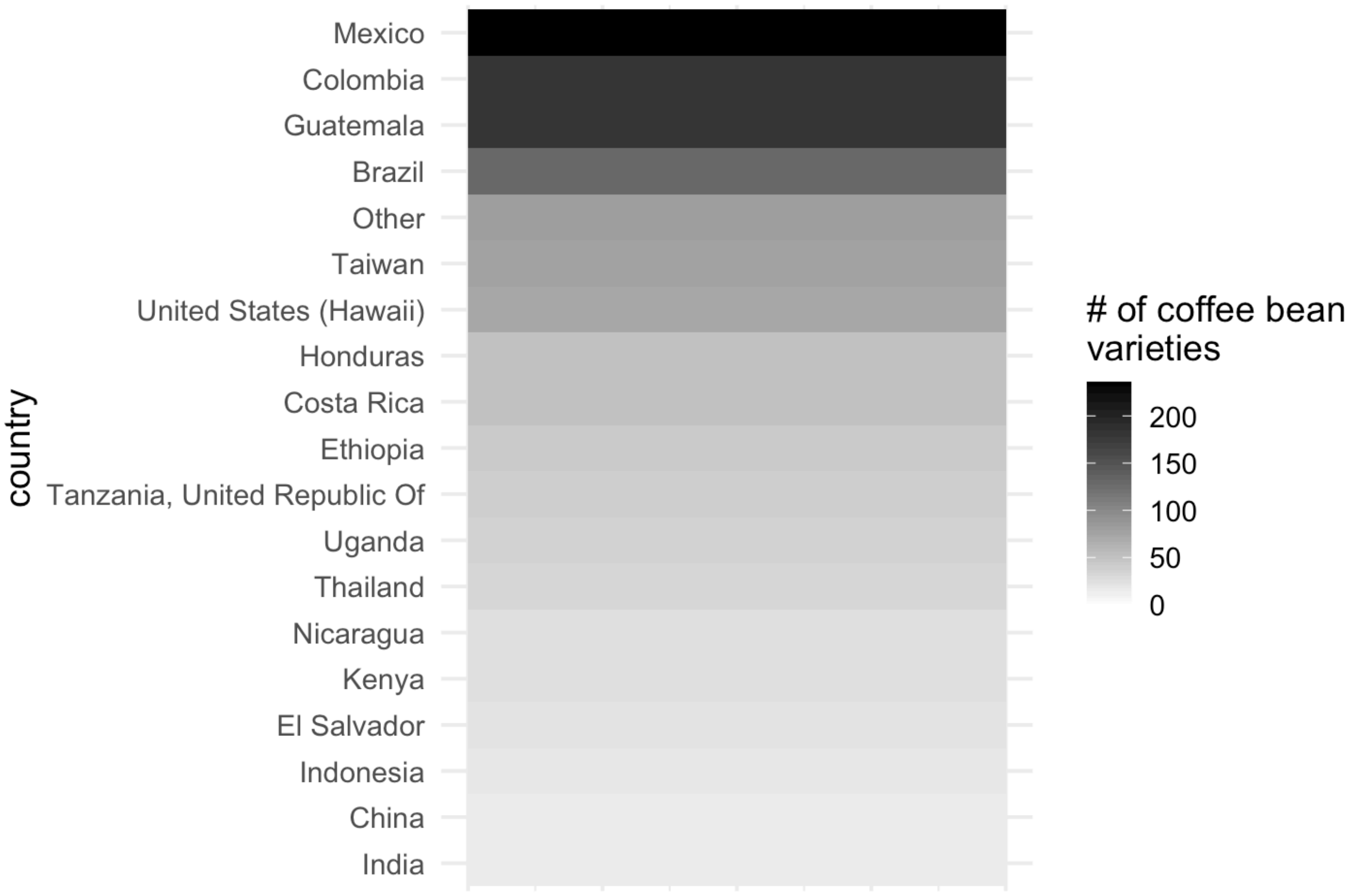


No legend?

No problem.

Because color saturation has natural ordering.

Color saturation is easier to quantify



The ratio between Mexico and United States is…

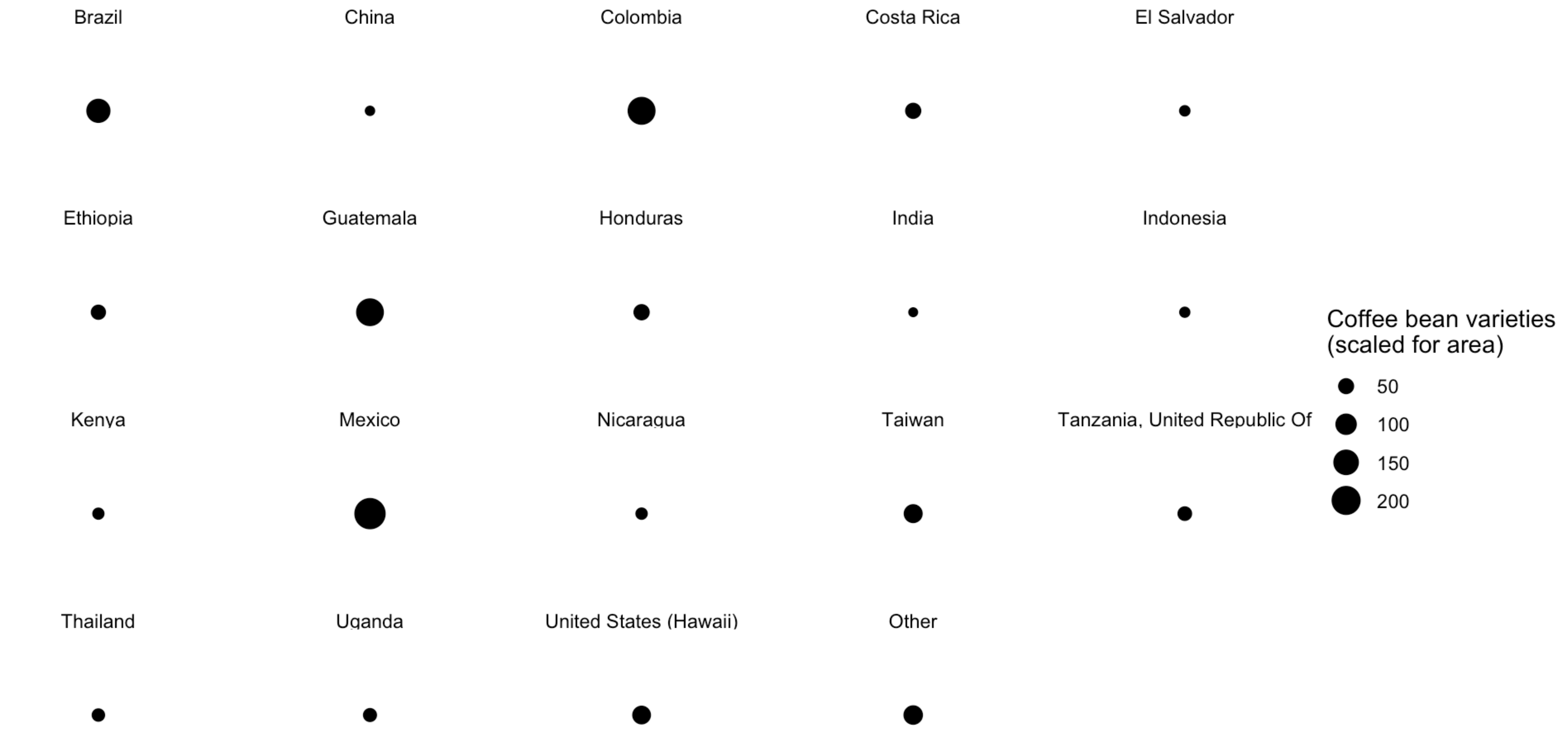
2 or 3

Moving down to the third

Move up one level to area

1. Position on a common scale
2. Position on non-aligned scales
3. Length
4. Angle
5. Area
6. Volume <> Density <> Color saturation
7. Color hue

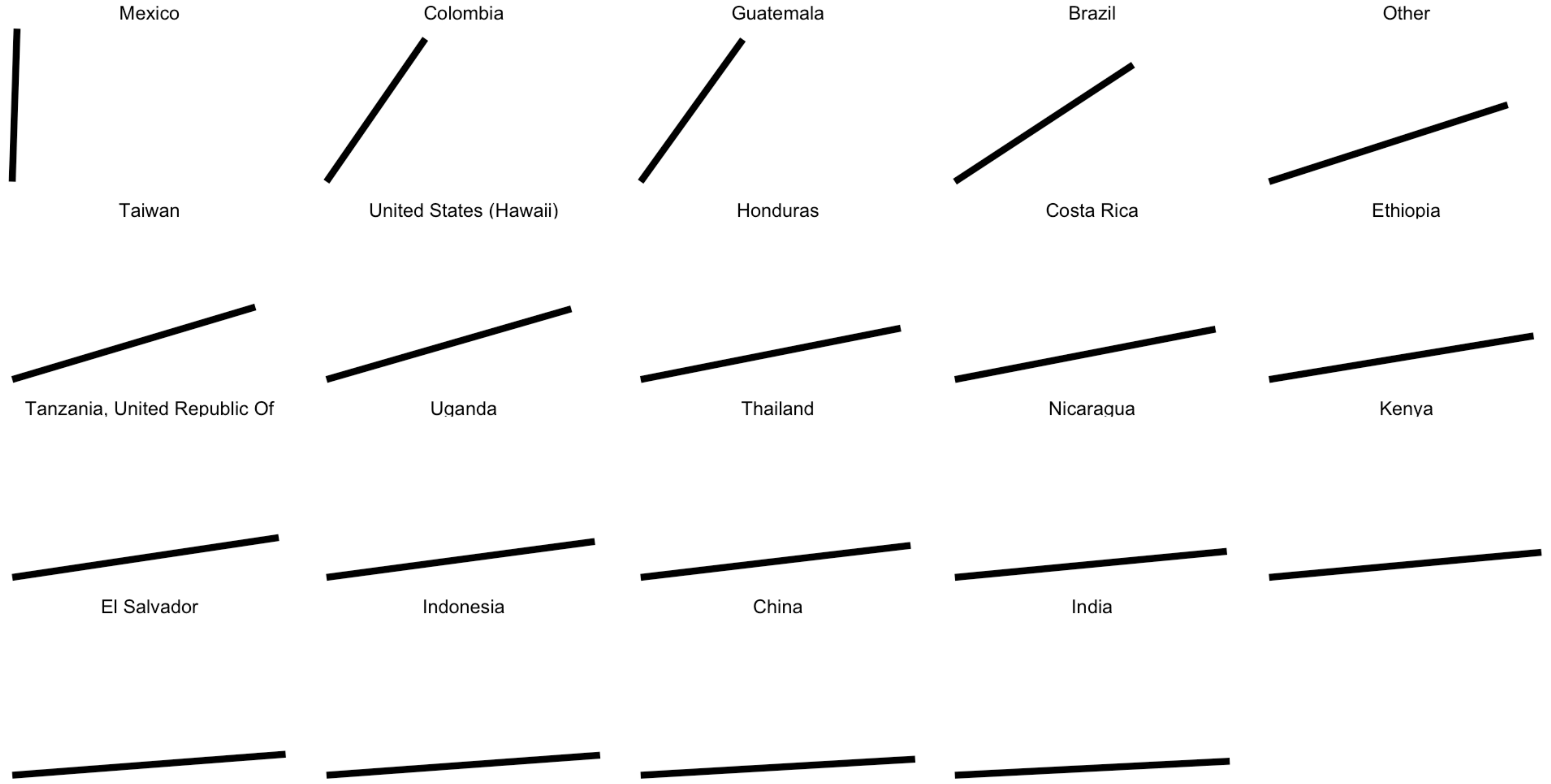
This is weird graph but still informative



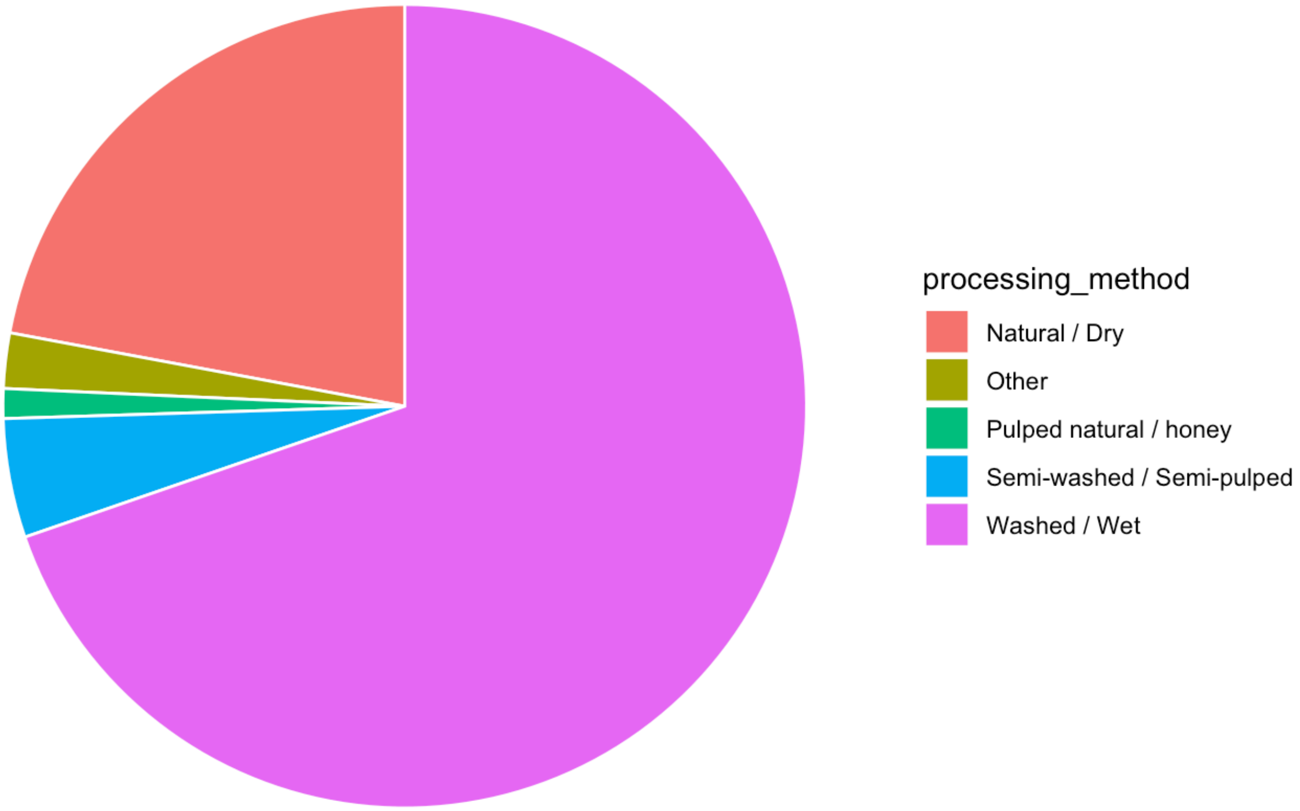
Move up one level to angle

1. Position on a common scale
2. Position on non-aligned scales
3. Length
4. Angle
5. Area
6. Volume <> Density <> Color saturation
7. Color hue

Use angle to visualize coffee bean varieties



Pie charts use angles to encode data

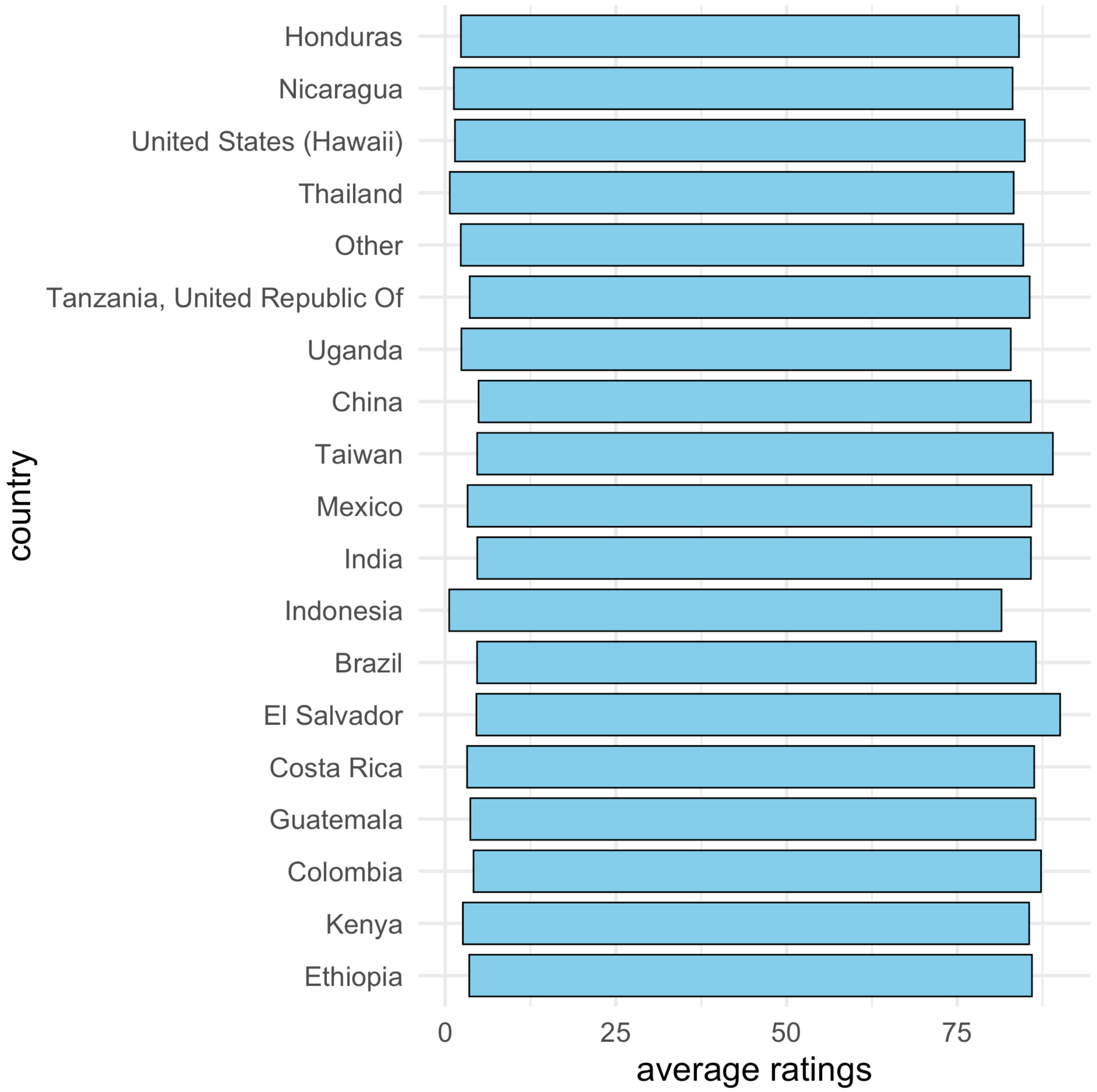


For categorical data, no more than 6 colors is best.

(Source: European Environment Agency)

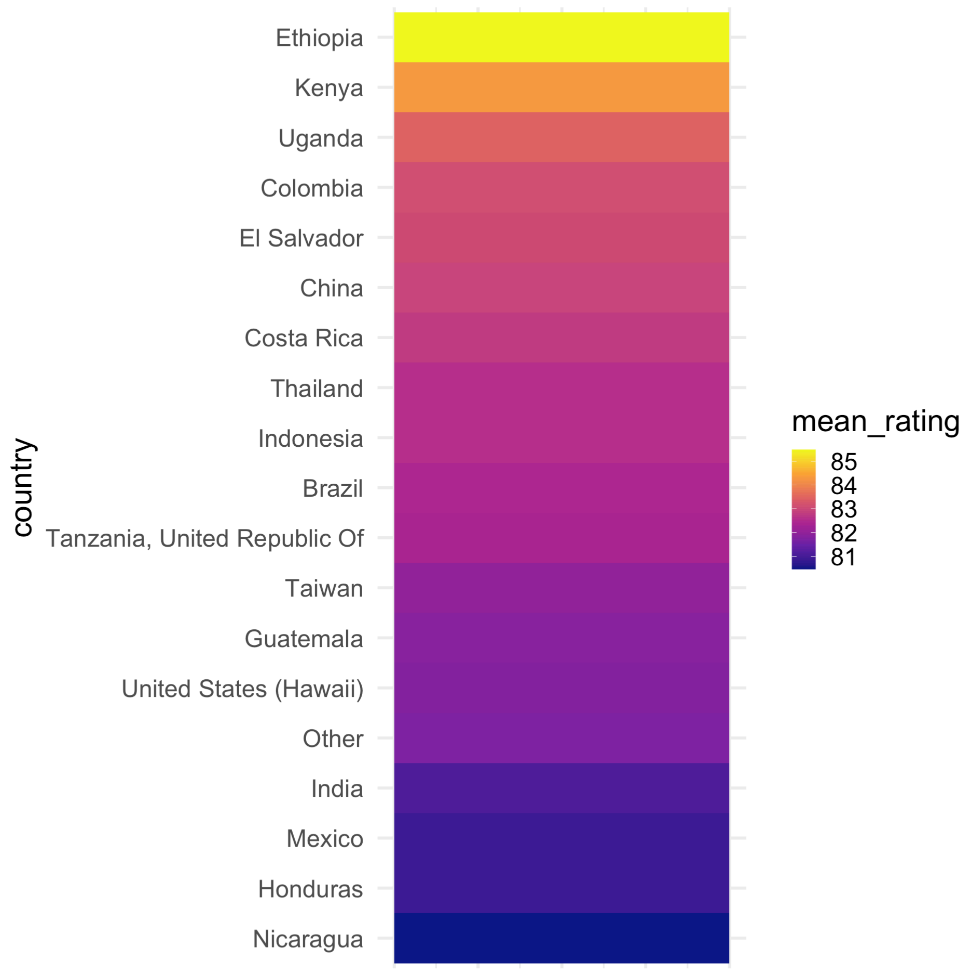
We are so close!

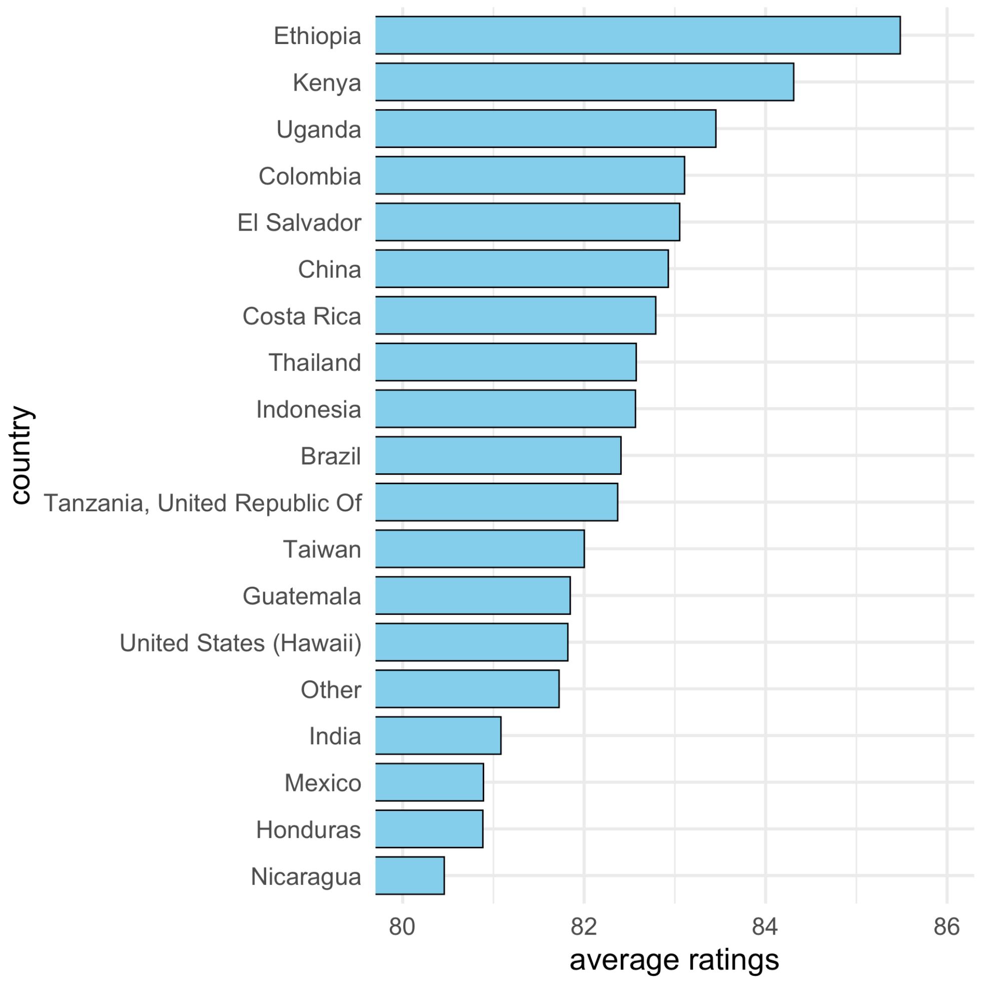
1. Position on a common scale
2. Position on non-aligned scales
3. Length
4. Angle
5. Area
6. Volume <> Density <> Color saturation
7. Color hue



Wait, I thought there is some difference…

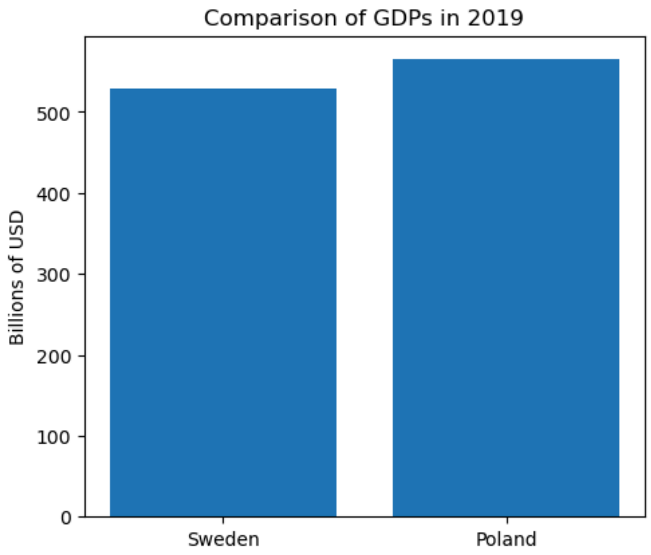
The start-at-zero rule



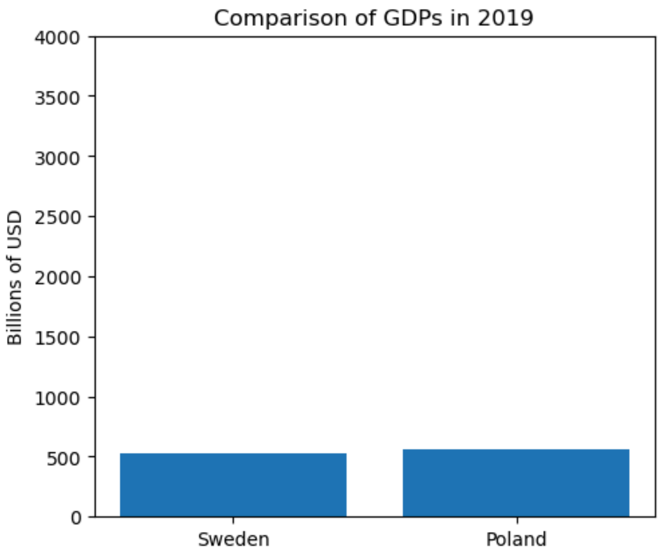


How to Lie with Statistics (1954)

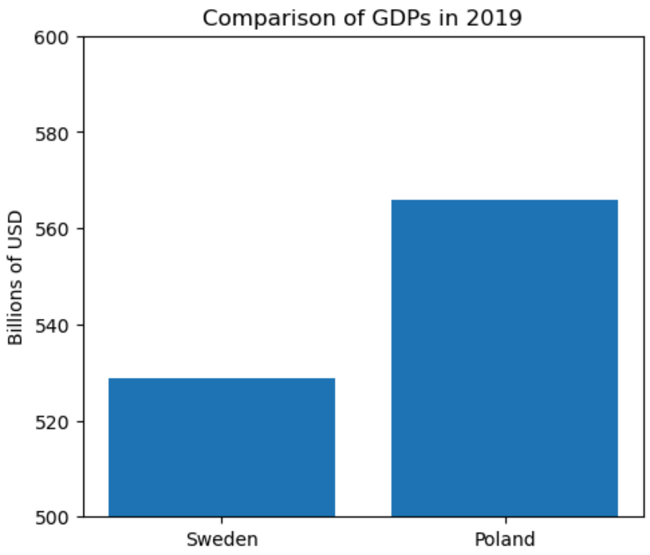
- Darrell Huff argues that truncating the y-axis can exaggerate differences and mislead the viewer.
- It creates a false impression of dramatic change where the actual variation is small.



Poland and Sweden are doing similarly great!



Oh, both have small GDPs...

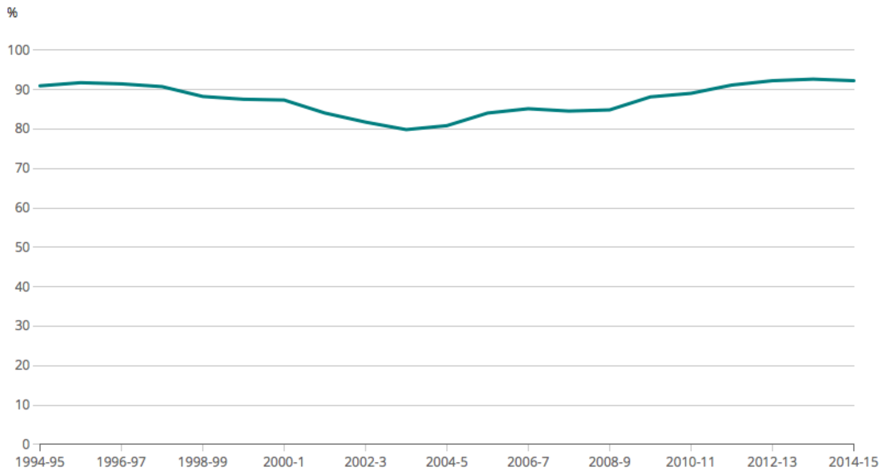


Poland is doing much better!

The Visual Display of Quantitative Information (1983)

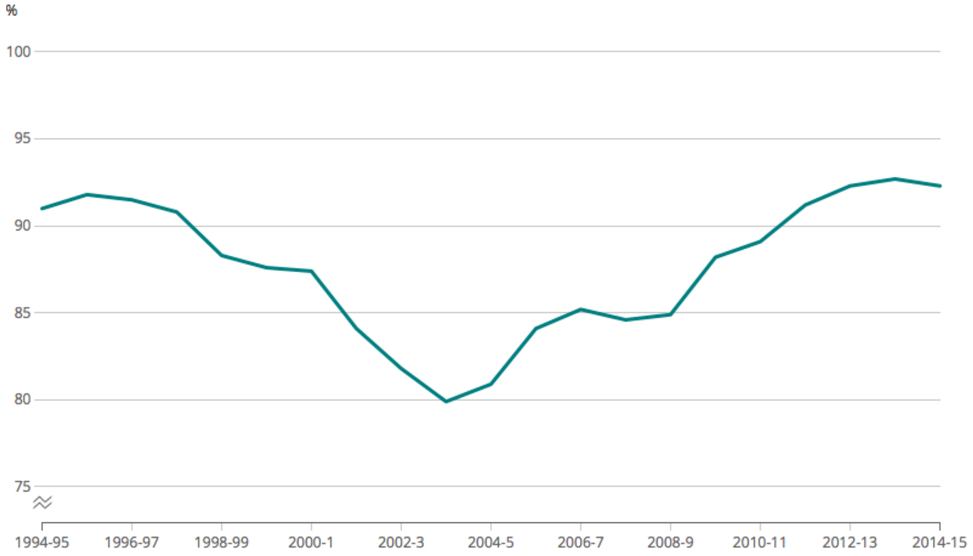
- Edward Tufte prioritizes data density and the detection of subtle patterns.
- He argues that starting at zero can waste valuable space, obscuring meaningful variations.

Combined MMR vaccination rate, 1994/95 to 2014/15, England



Source: NHS Immunisation Statistics - England, 2014-15, Table 8 and 9, HSCIC

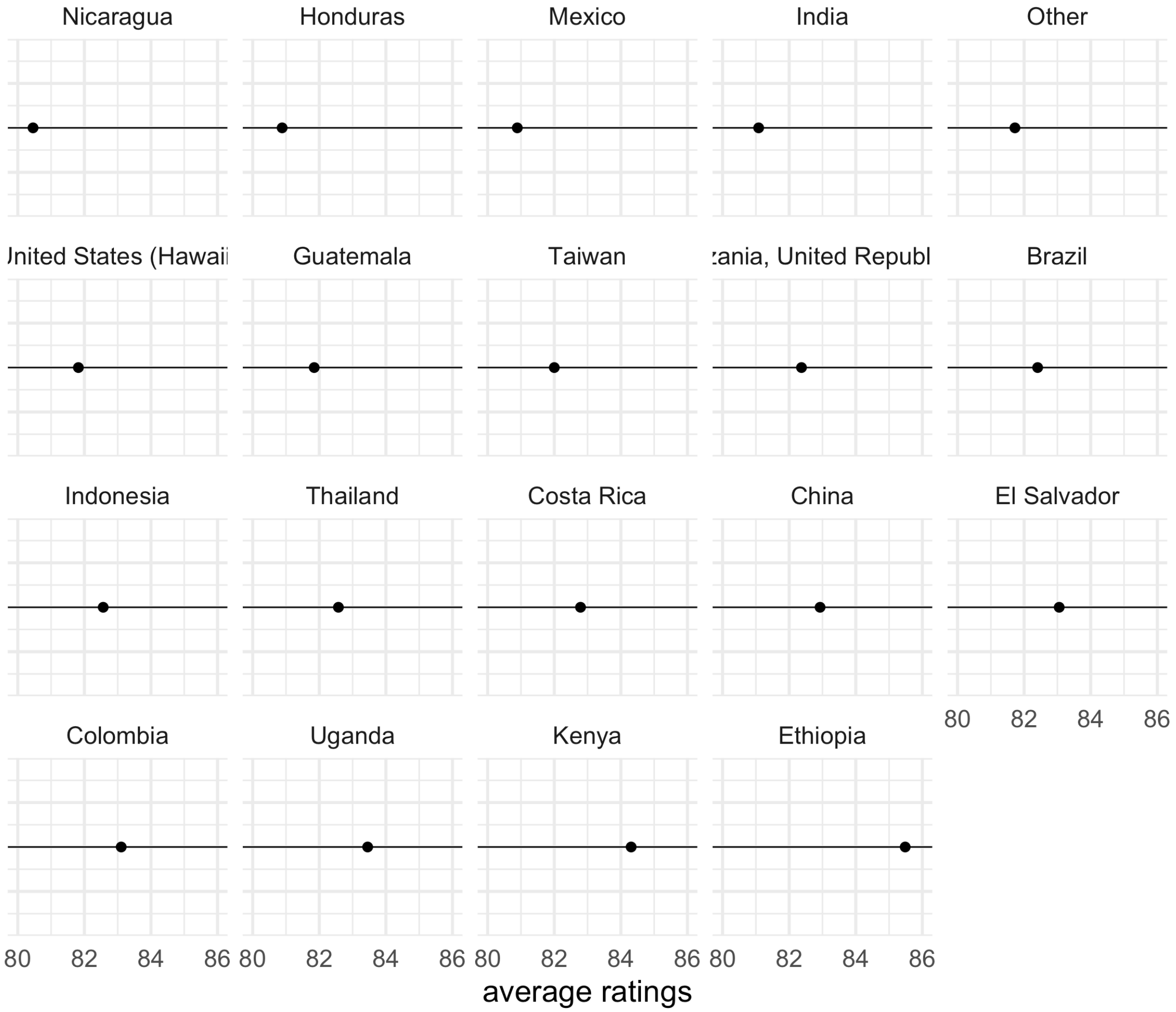
Take another look, axis doesn't start at zero



Source: NHS Immunisation Statistics - England, 2014-15, Table 8 and 9, HSCIC

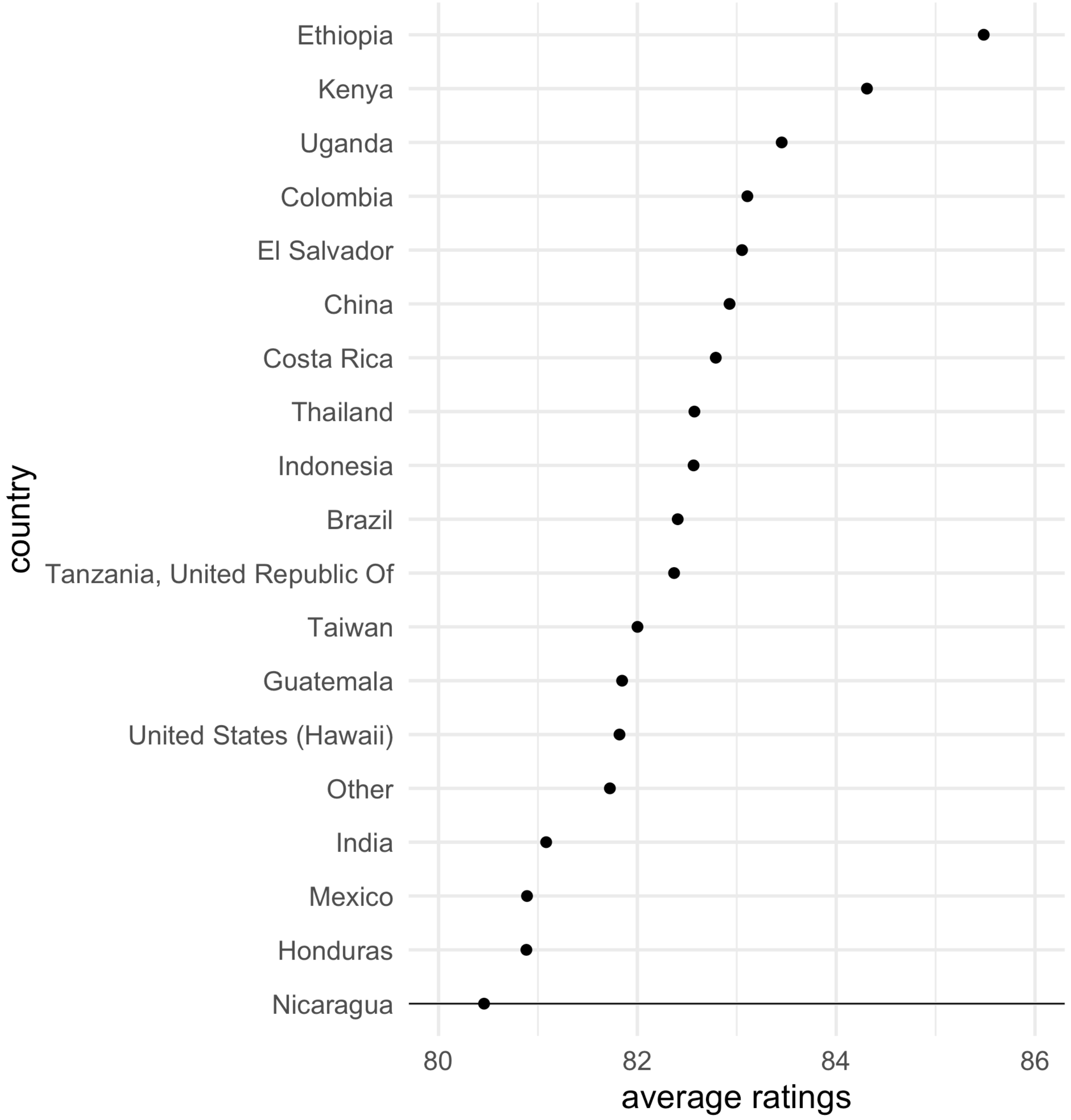
Position, but not a common scale

1. Position on a common scale
2. Position on non-aligned scales
3. Length
4. Angle
5. Area
6. Volume <> Density <> Color saturation
7. Color hue



Position, and a common scale

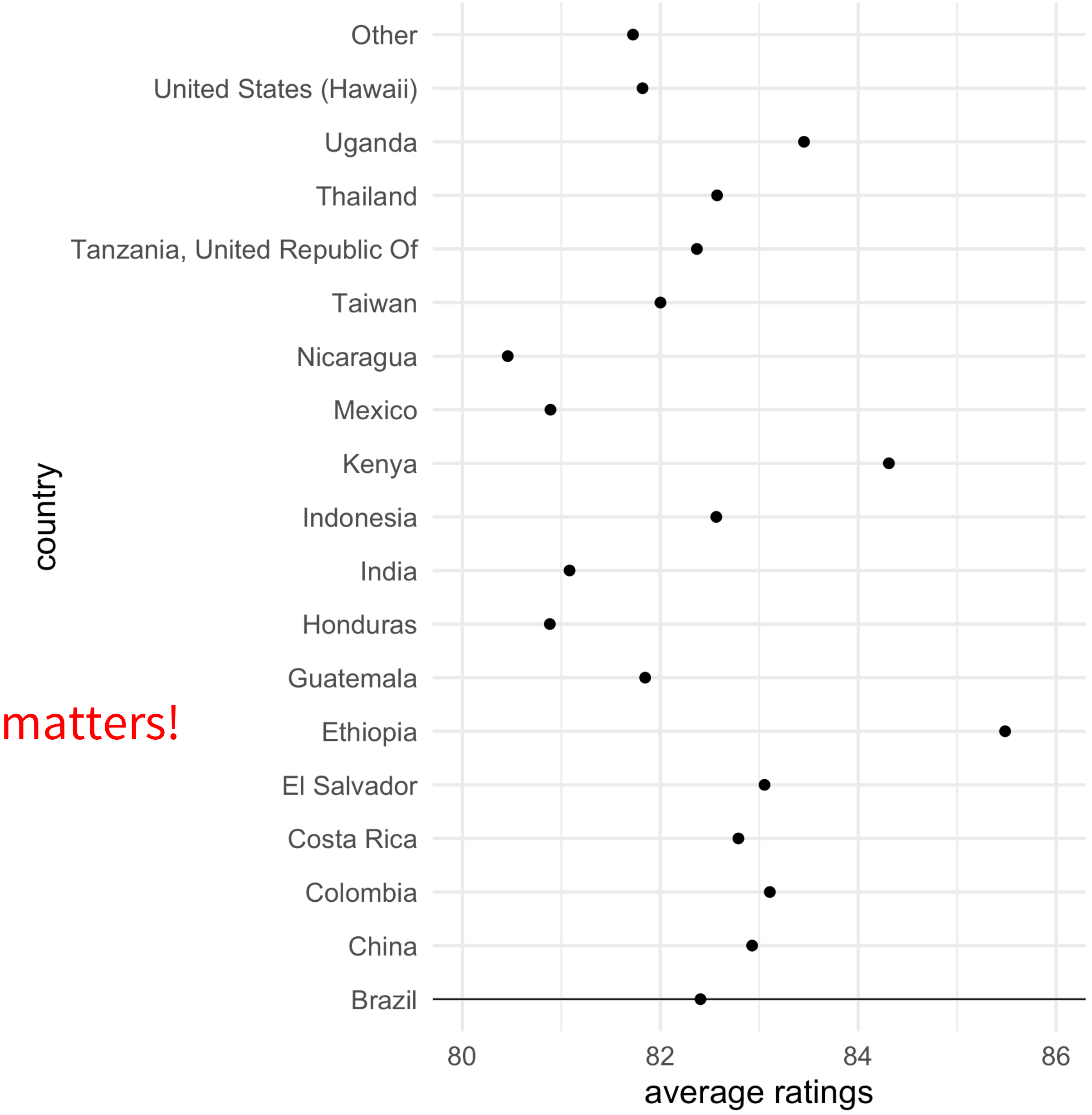
1. Position on a common scale
2. Position on non-aligned scales
3. Length
4. Angle
5. Area
6. Volume <> Density <> Color saturation
7. Color hue



Position, and a common scale

1. Position on a common scale
2. Position on non-aligned scales
3. Length
4. Angle
5. Area
6. Volume <> Density <> Color saturation
7. Color hue

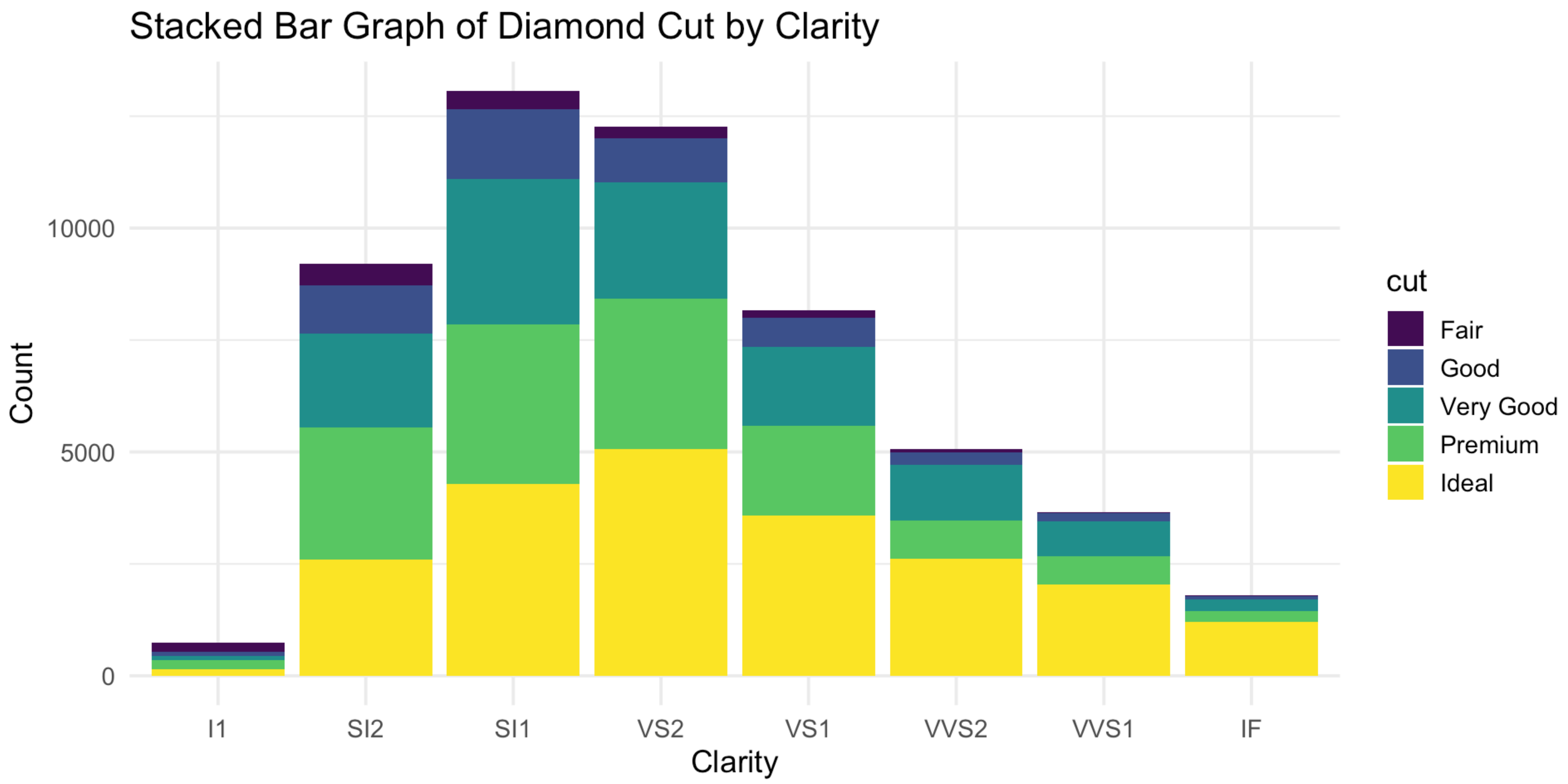
Re-ranking categorical variables still matters!



Implications for designing effective data visualizations

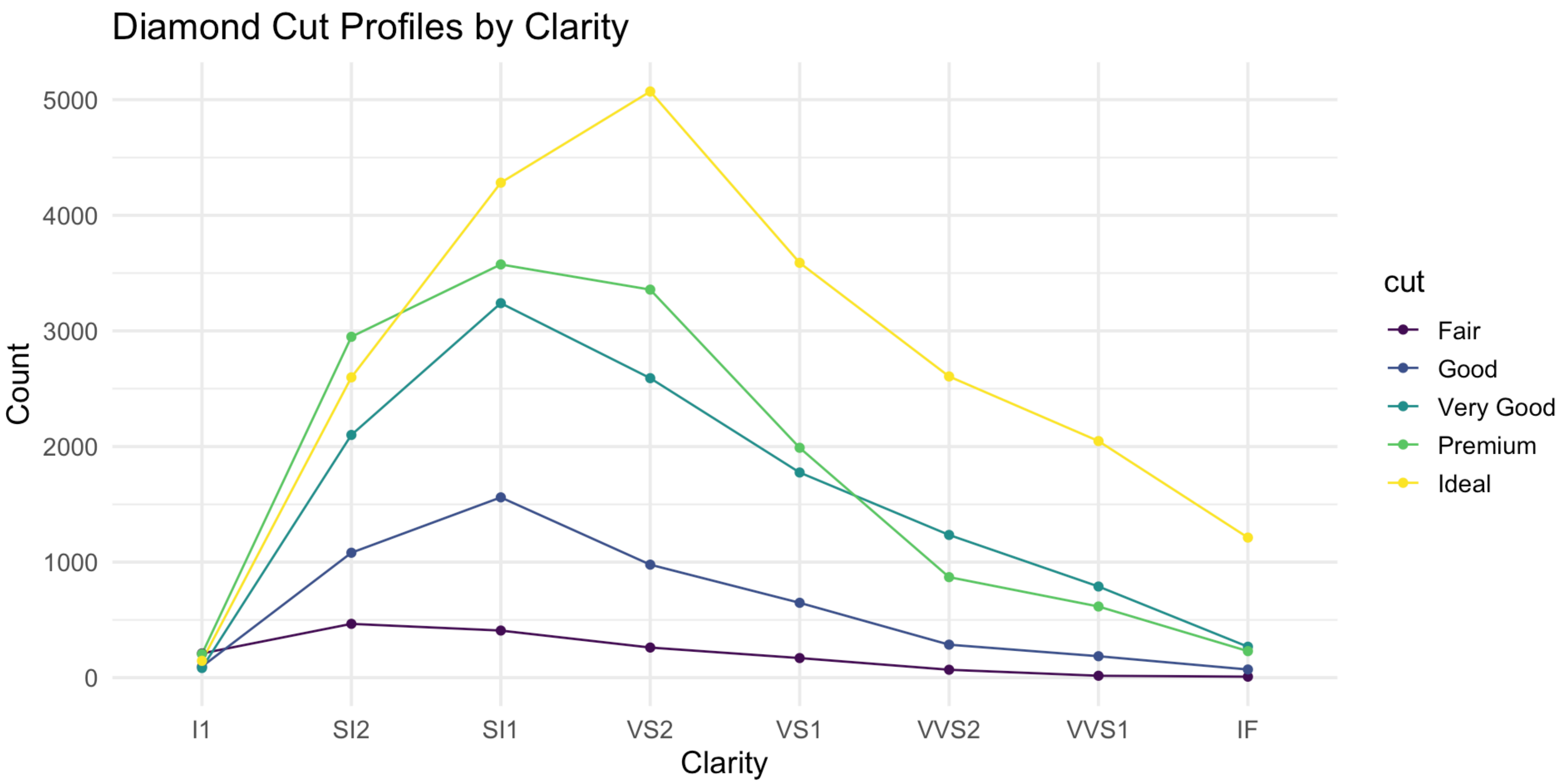
- Stacked anything is nearly always a mistake
- Pie charts are always a mistake
- Scatterplot are the best way to show the relationships between two variables
- If growth (slope) is important, plot it directly

Stacked anything is nearly always a mistake!



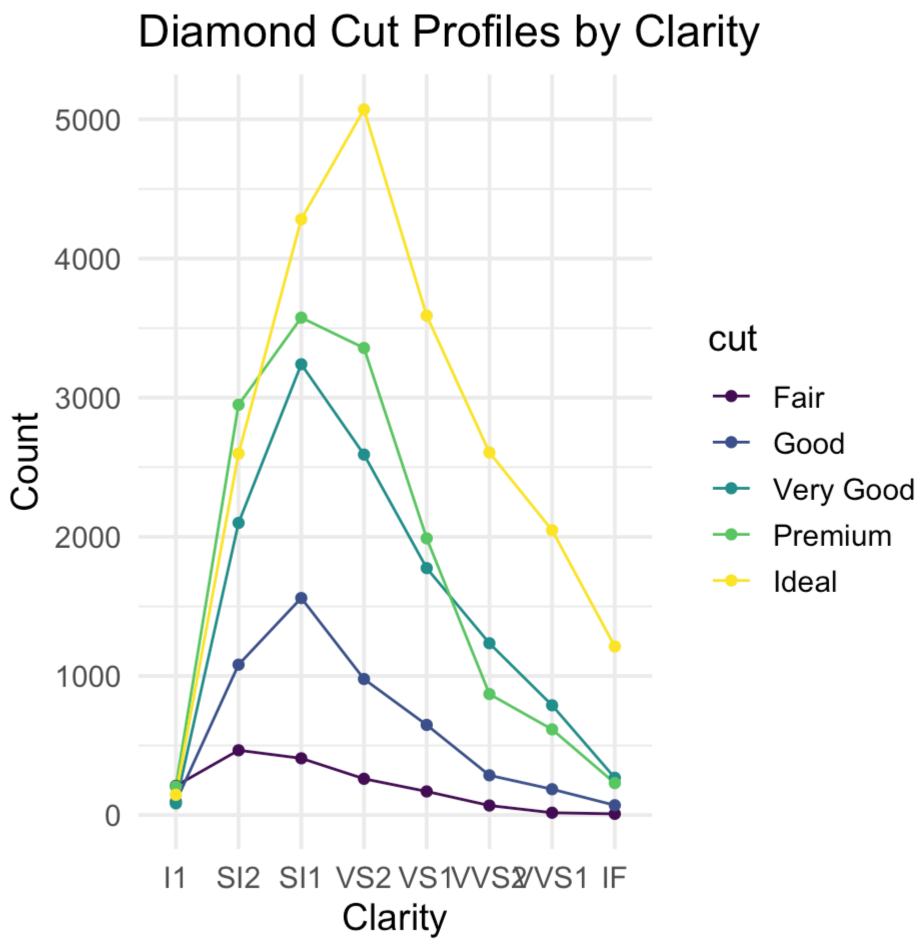
Which category has higher count: SI1-Premium or VS2-Premium?

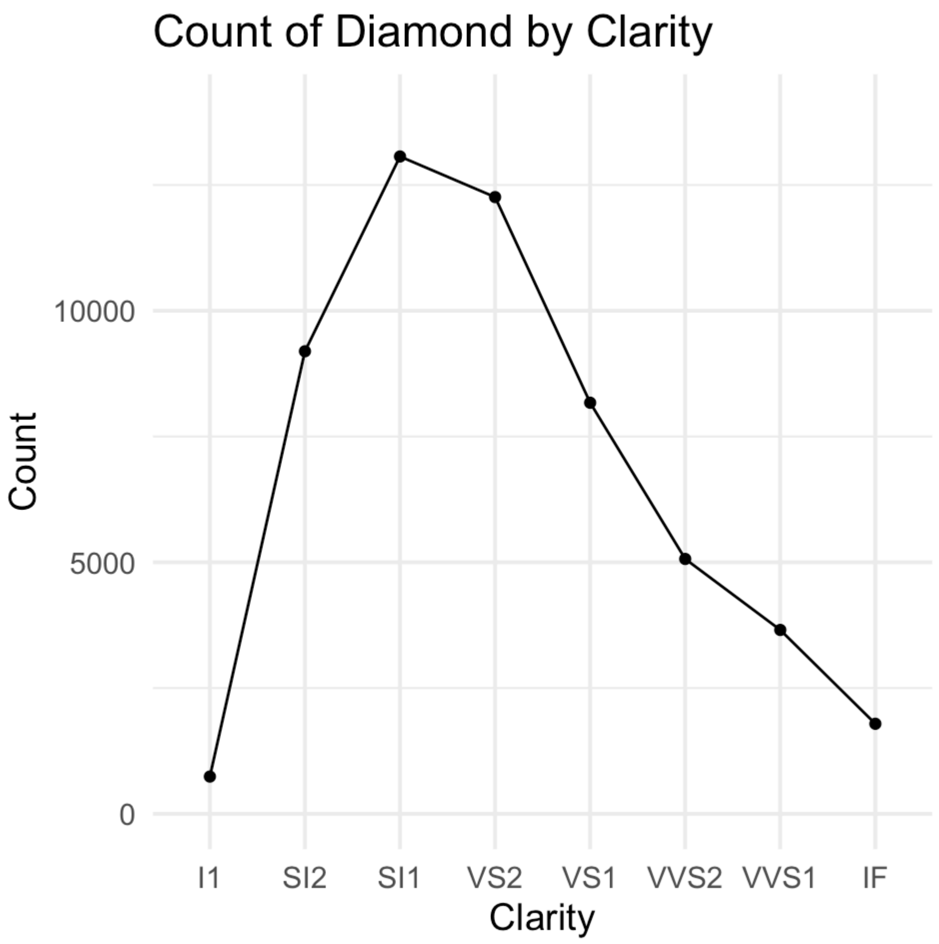
Transform stacked barplot to a parallel coordinate plot



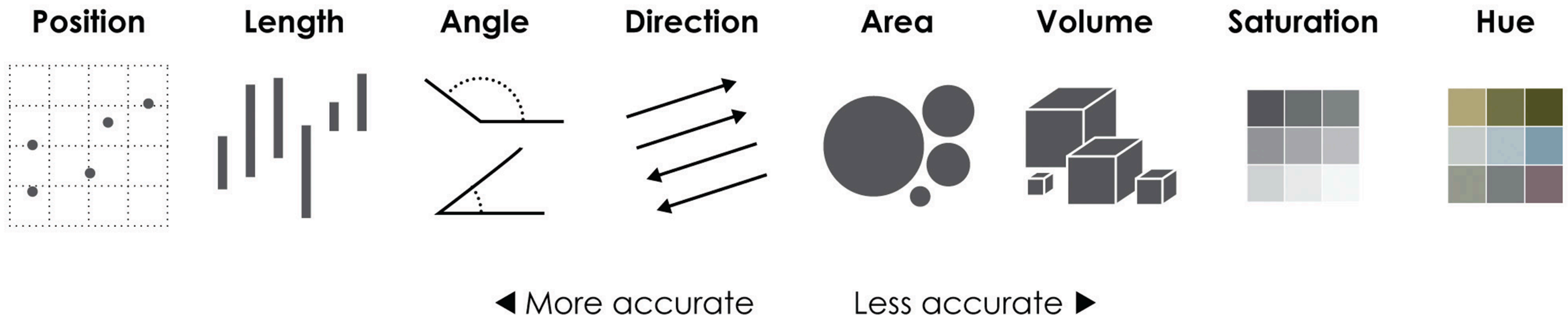
Which category has higher count: SI1-Premium or VS2-Premium?

You lose some information, but just use two charts if needed





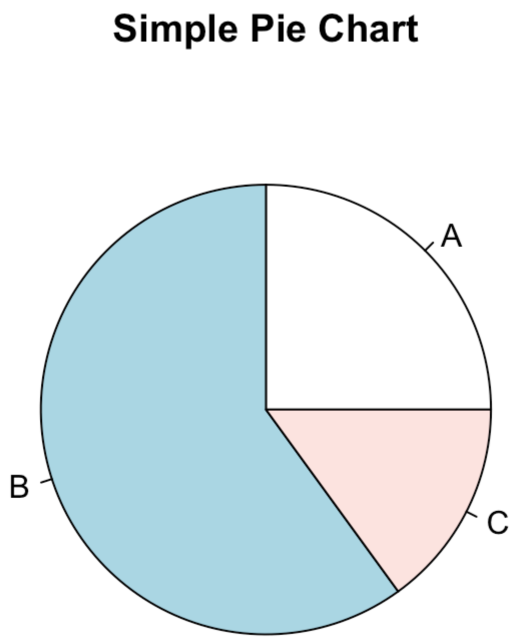
Why are pie charts never a good idea?



Angle is #4 on the accuracy list, we can do better.

If you have a small amount of data to show, don’t use pie charts

Don’t do this!



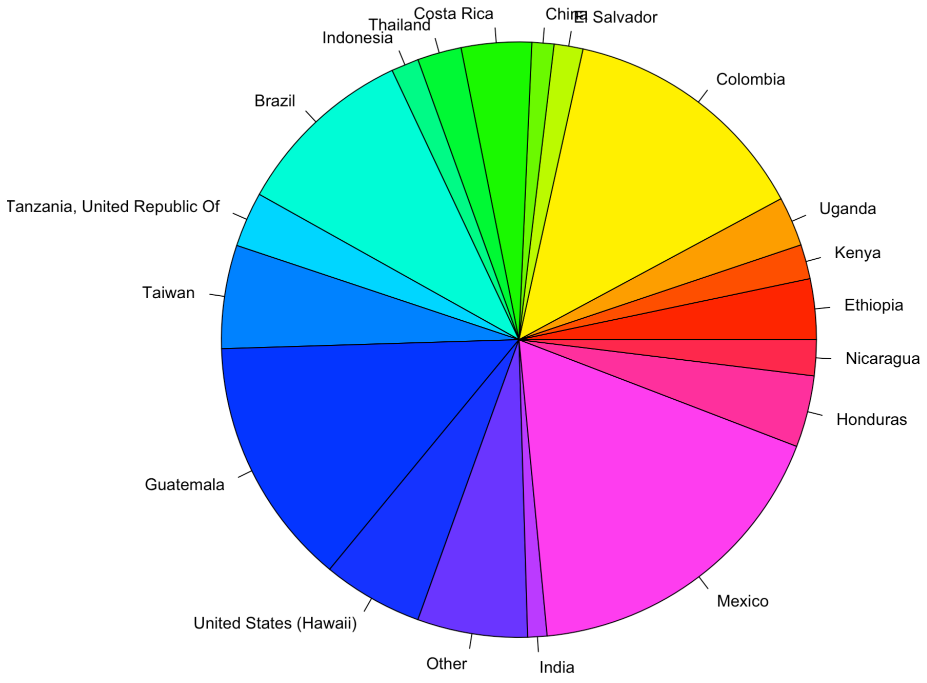
Do this instead!

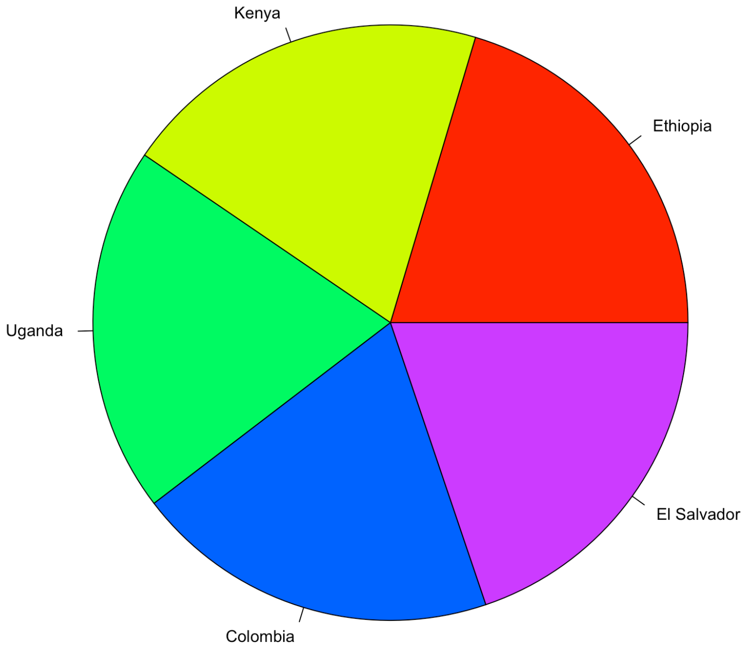
Label	Value
A	25
B	60
C	15

If you have a lot of data to show, don’t use pie charts

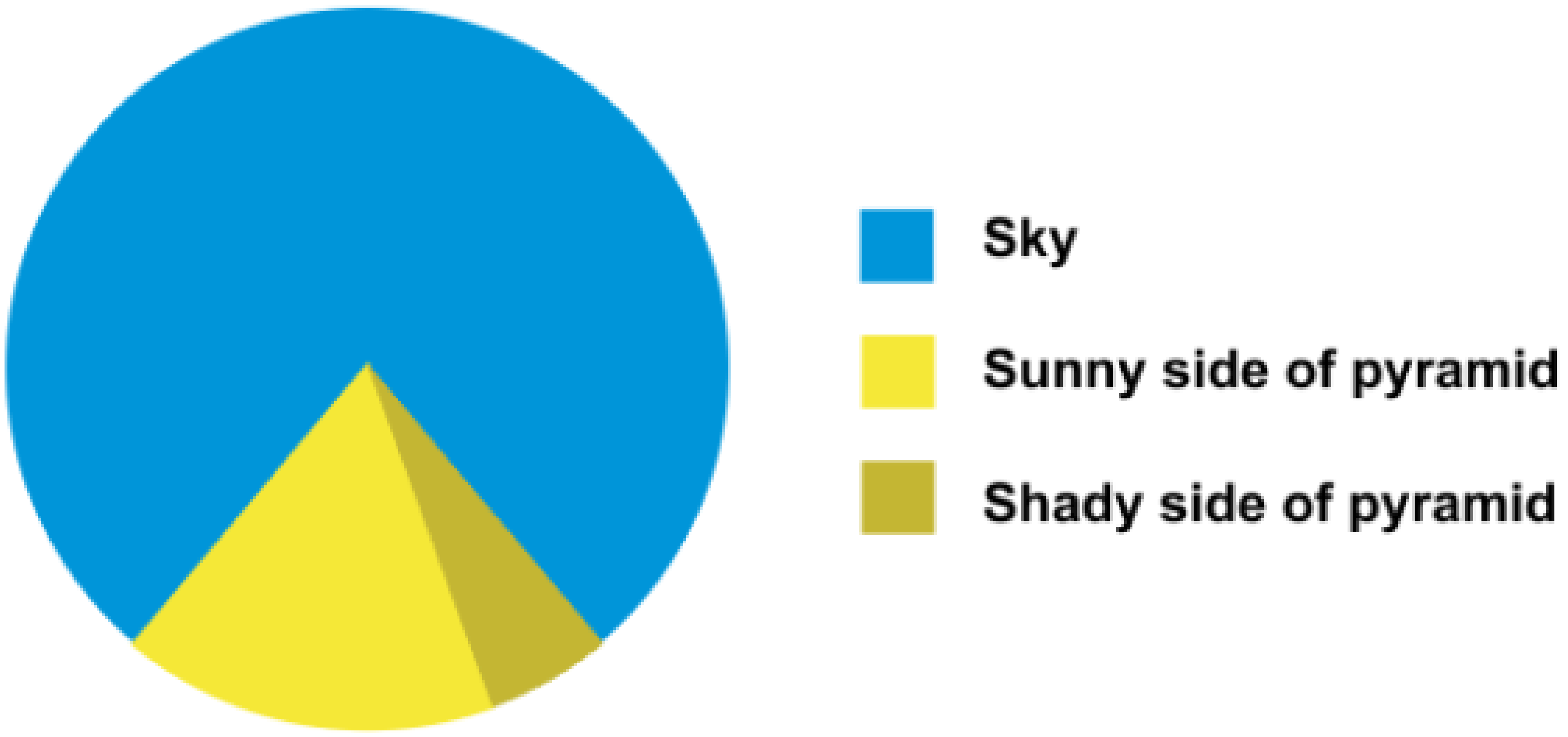
Don’t do this!

Or this!

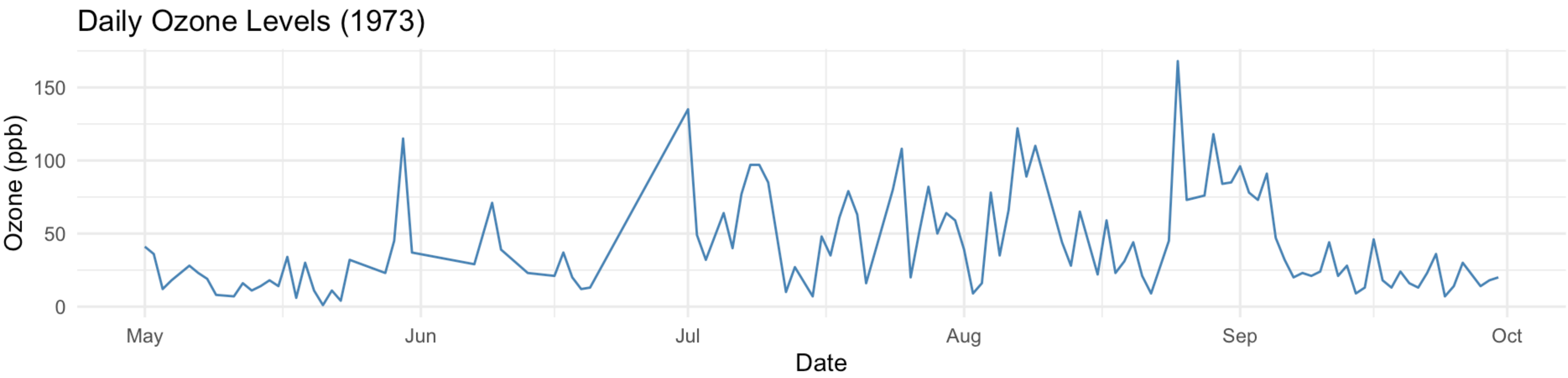


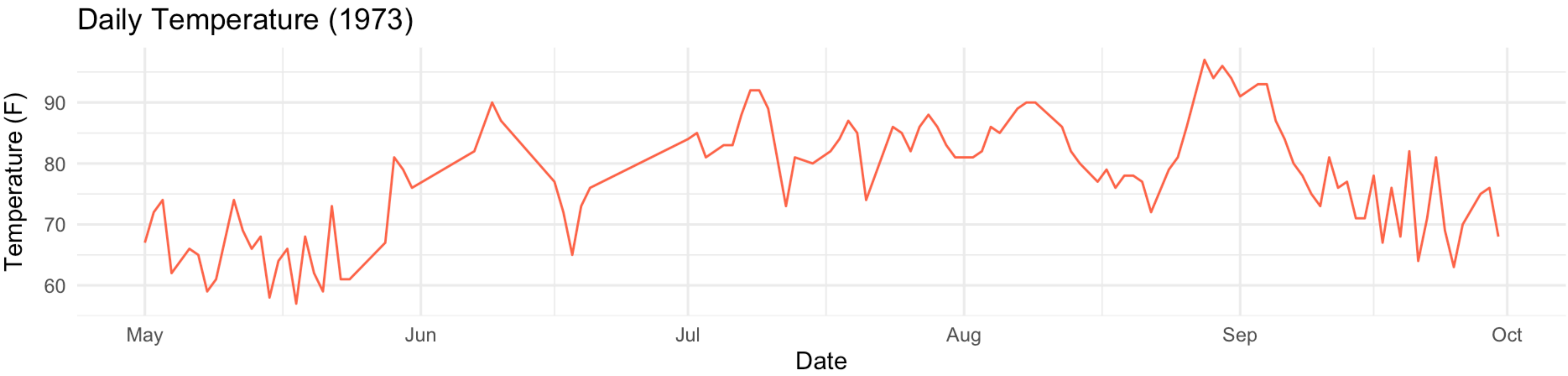


All good pie charts are jokes



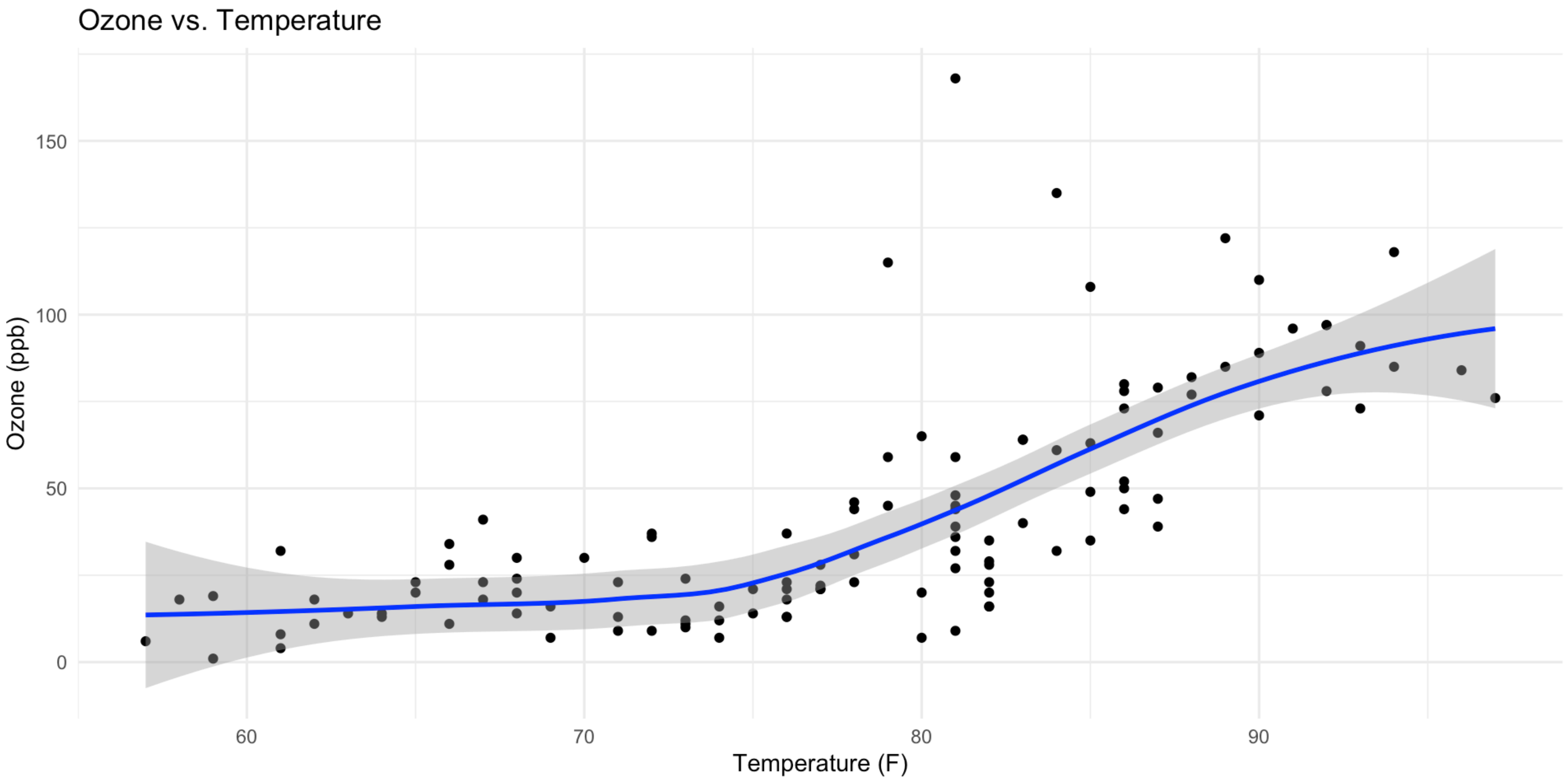
If you want to show the relationship between two variables, use scatterplot


Daily Ozone Levels (1973)

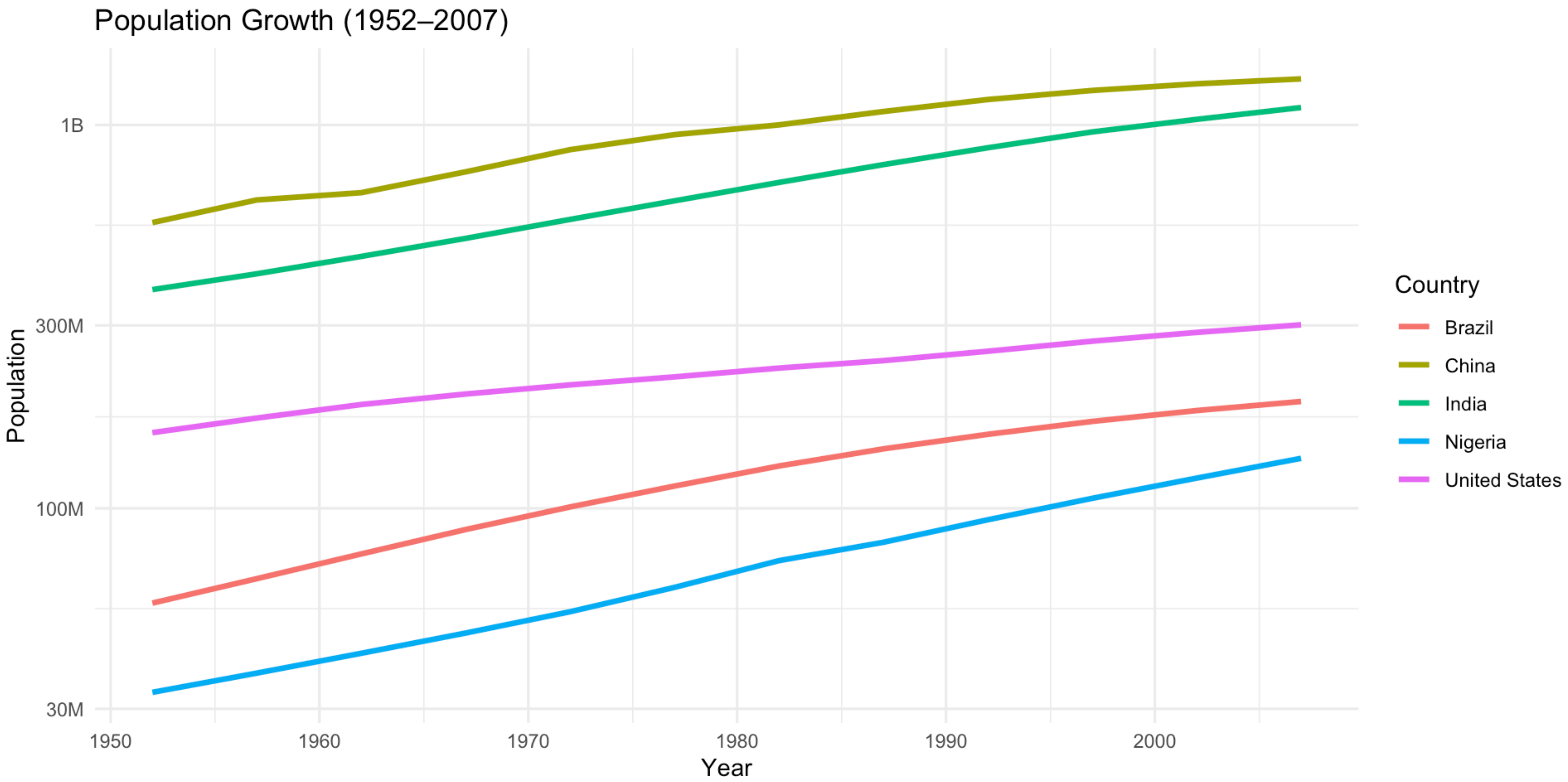

Daily Temperature (1973)

What is the relationship between Ozone concentrations and temperature?

If you want to show the relationship between two variables, use scatterplot

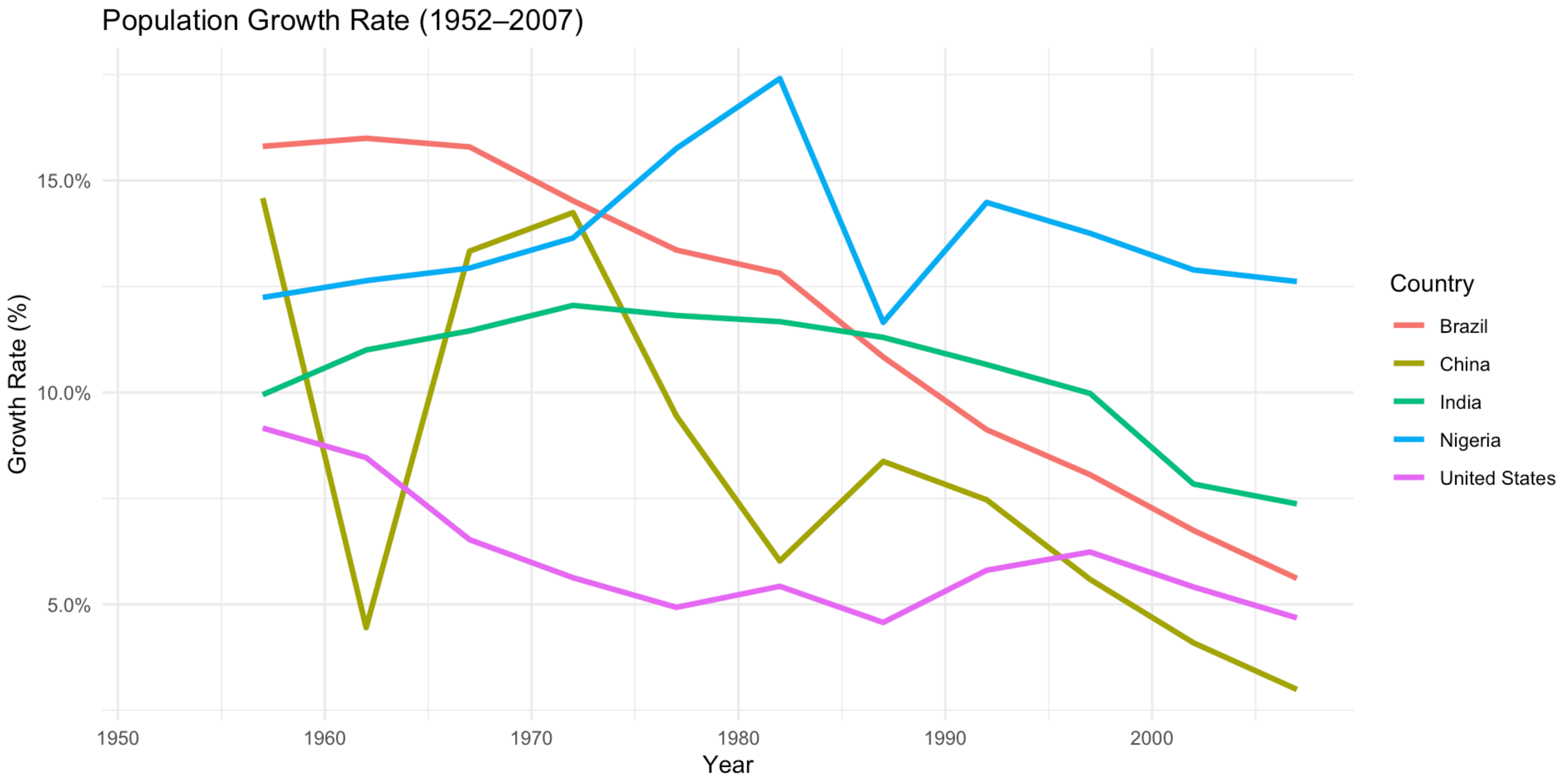


If you care about the growth (slope), plot it directly



Which country has higher population growth: Nigeria or India?

If you care about the growth (slope), plot it directly



Cleveland’s three visual operations of pattern perception

🎯 **Detection**: *Recognizing that a geometric object encodes a physical value.*

🧩 **Assembly**: *Grouping detected graphical elements into patterns.*

📏 **Estimation**: *Visually assessing the relative magnitude of two or more values.*

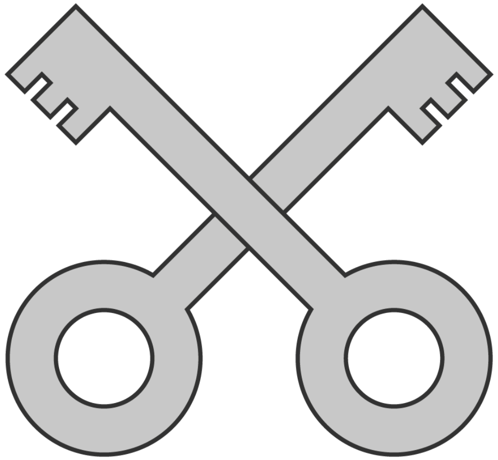
Assembly: Gestalt Psychology

> “Gestalt (German for form, shape, or configuration). Gestalt psychology proposes that the human brain perceives objects as part of a greater whole rather than as isolated elements.”

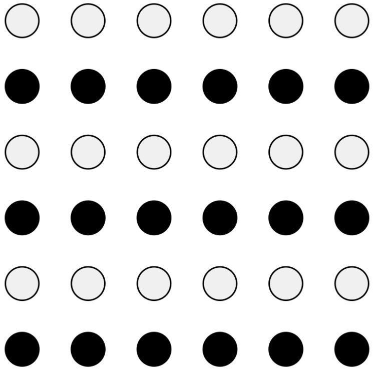
Applying Gestalt principles to data visualization

> “The law of **Prãgnanz**, also known as the law of good Gestalt. People tend to experience things as regular, orderly, symmetrical, and simple.”

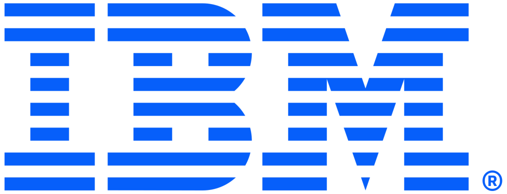
Law of Continuity



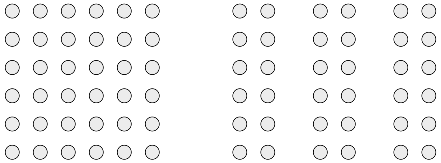
Law of Similarity



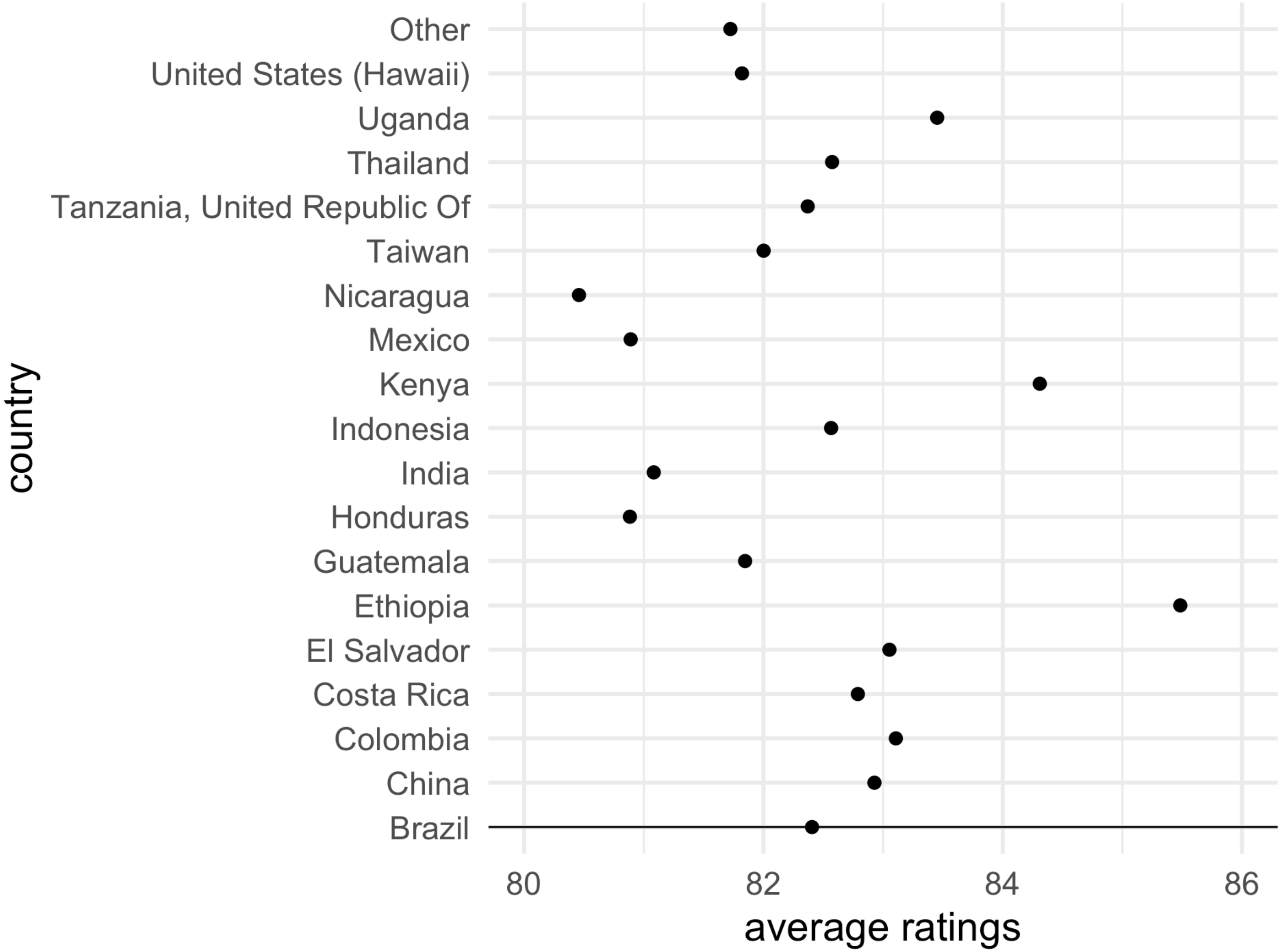
Law of Closure



Law of Proximity

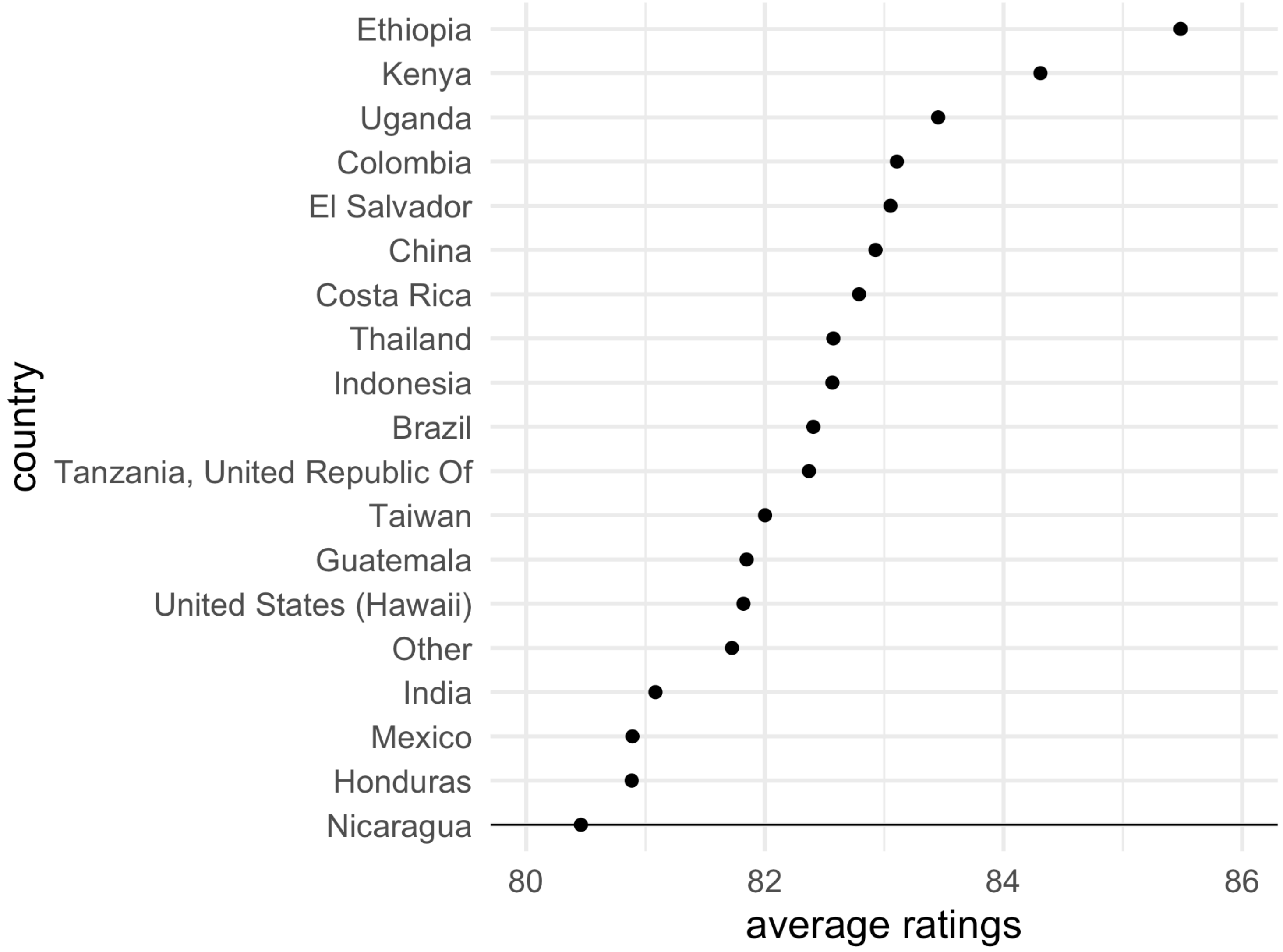


Bad visualizations lack law of continuity



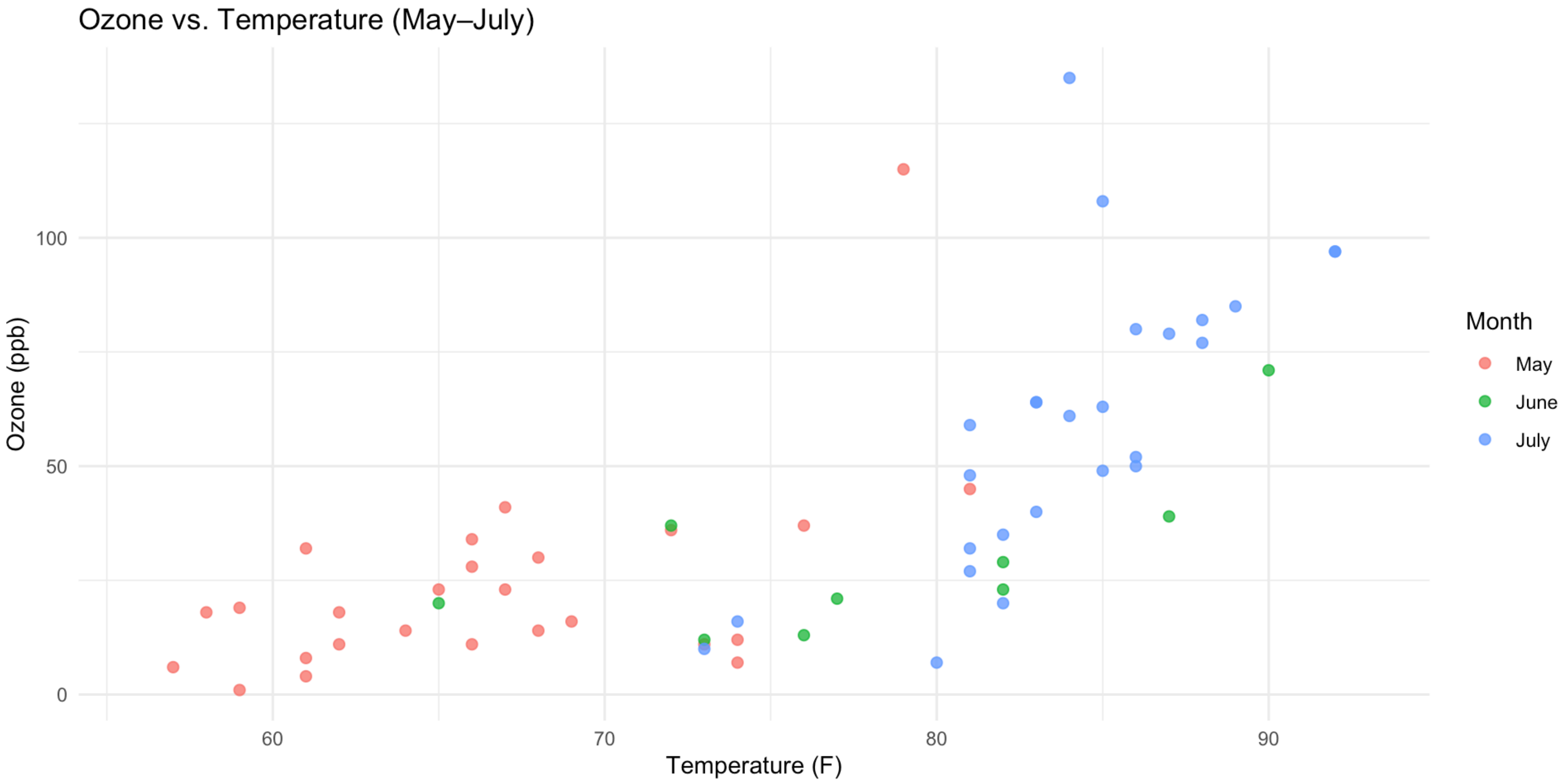
This hurts our brain.

Good visualizations leverage law of continuity

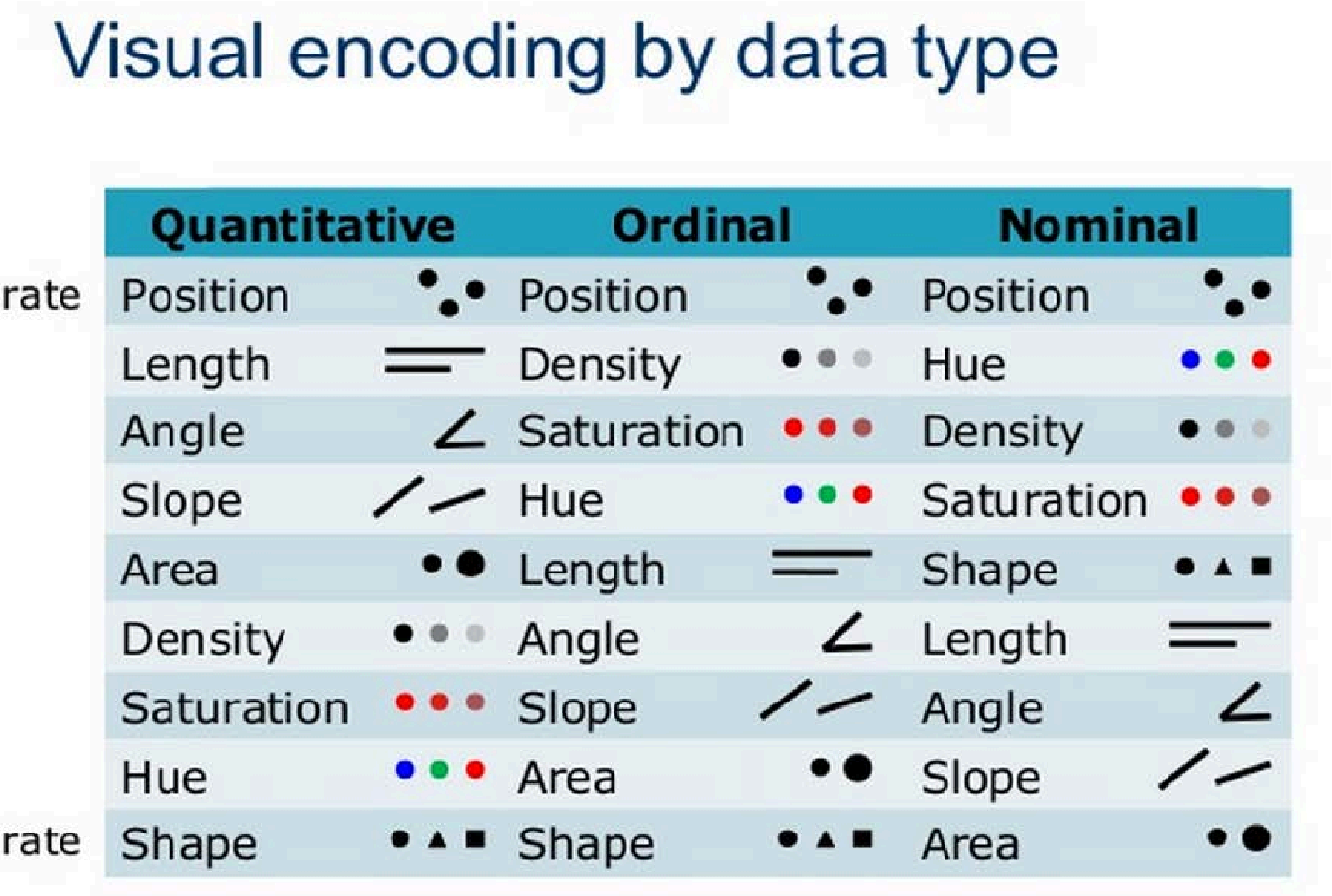


This is much easier.

Use law of similarity to group similar data



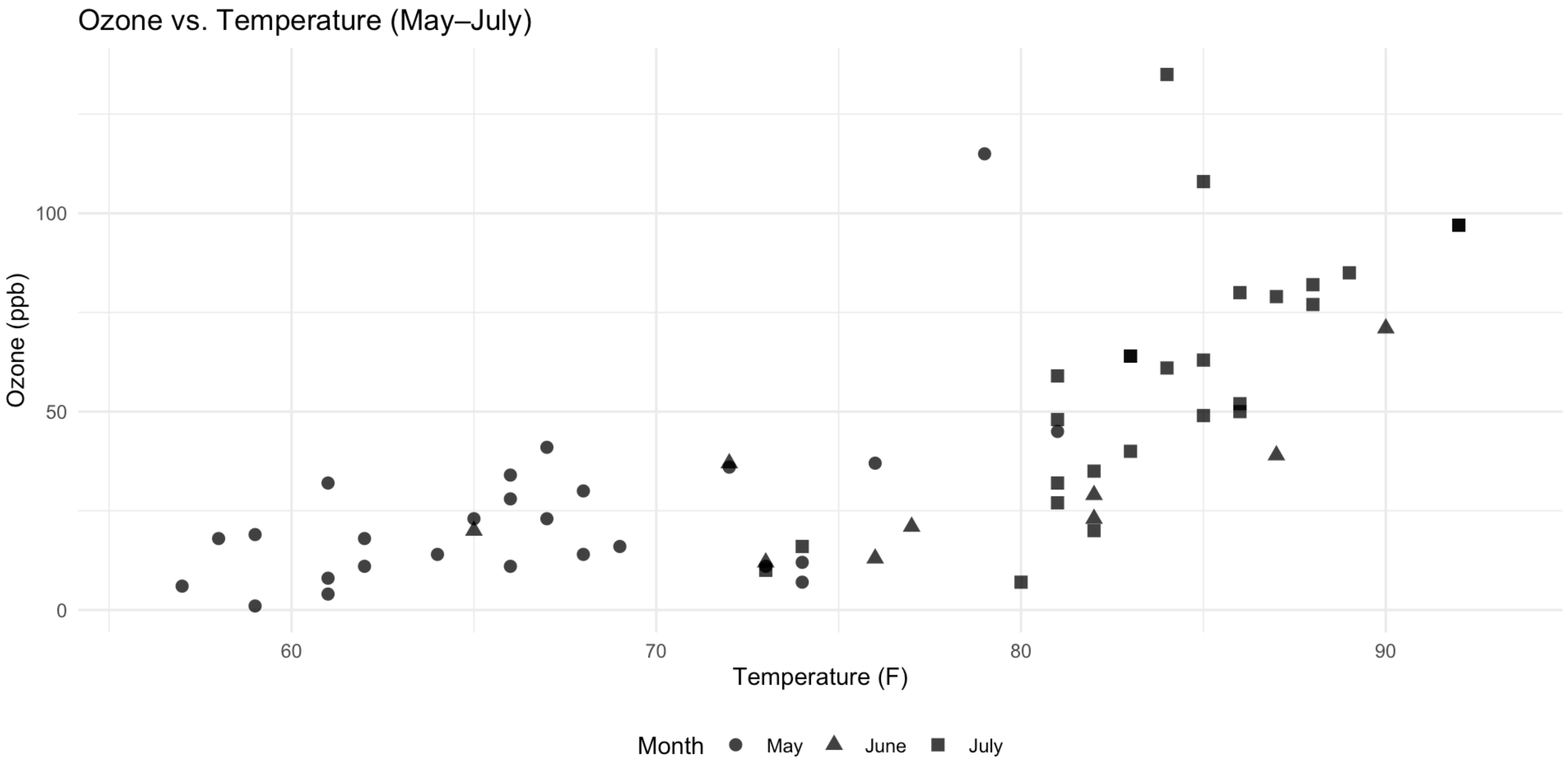
Some encodings are better than others



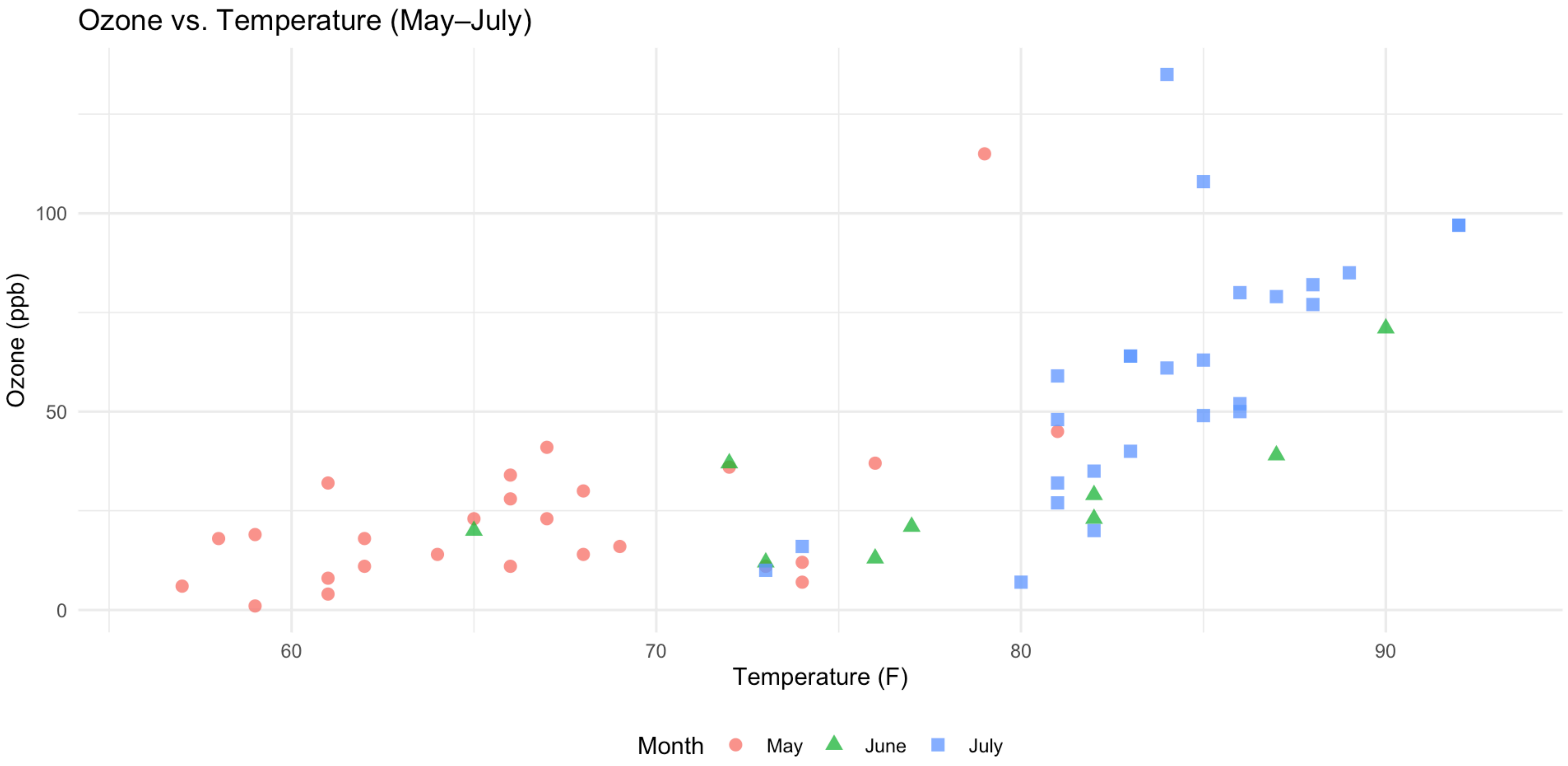
Visual encoding by data type

	Quantitative		Ordinal		Nominal	
More Accurate	Position	• • •	Position	• • •	Position	• • •
	Length	=	Density	● ● ●	Hue	🔵 🟢 🔴
	Angle	∠	Saturation	🔴 🔴 ●	Density	● ● ●
	Slope	/ /	Hue	🔵 🟢 🔴	Saturation	🔴 🔴 ●
	Area	• ●	Length	=	Shape	● ▲ ■
	Density	● ● ●	Angle	∠	Length	=
	Saturation	🔴 🔴 ●	Slope	/ /	Angle	∠
	Hue	🔵 🟢 🔴	Area	• ●	Slope	/ /
Less Accurate	Shape	● ▲ ■	Shape	● ▲ ■	Area	• ●

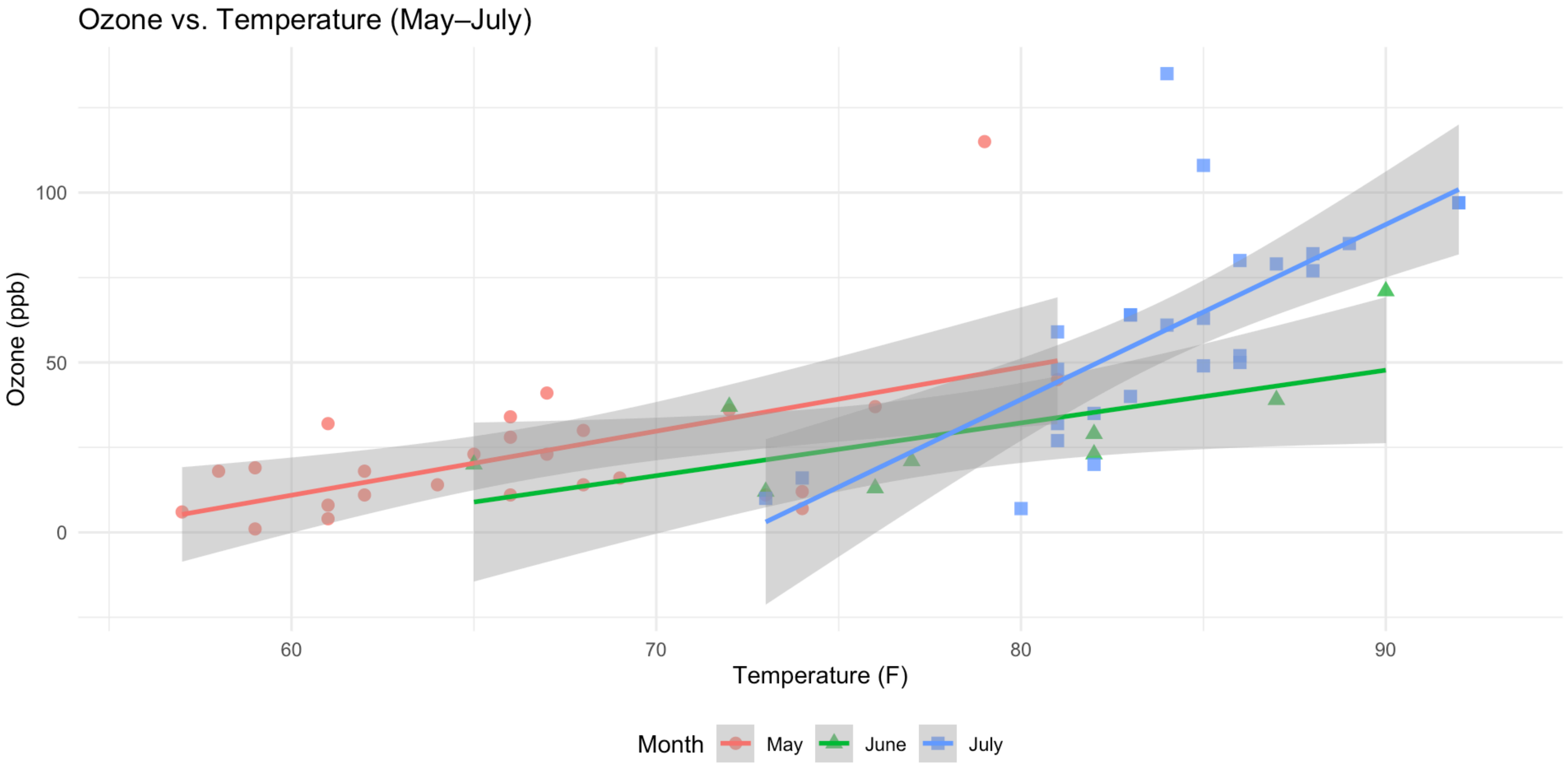
Shape is less effective than color hue for nominal data



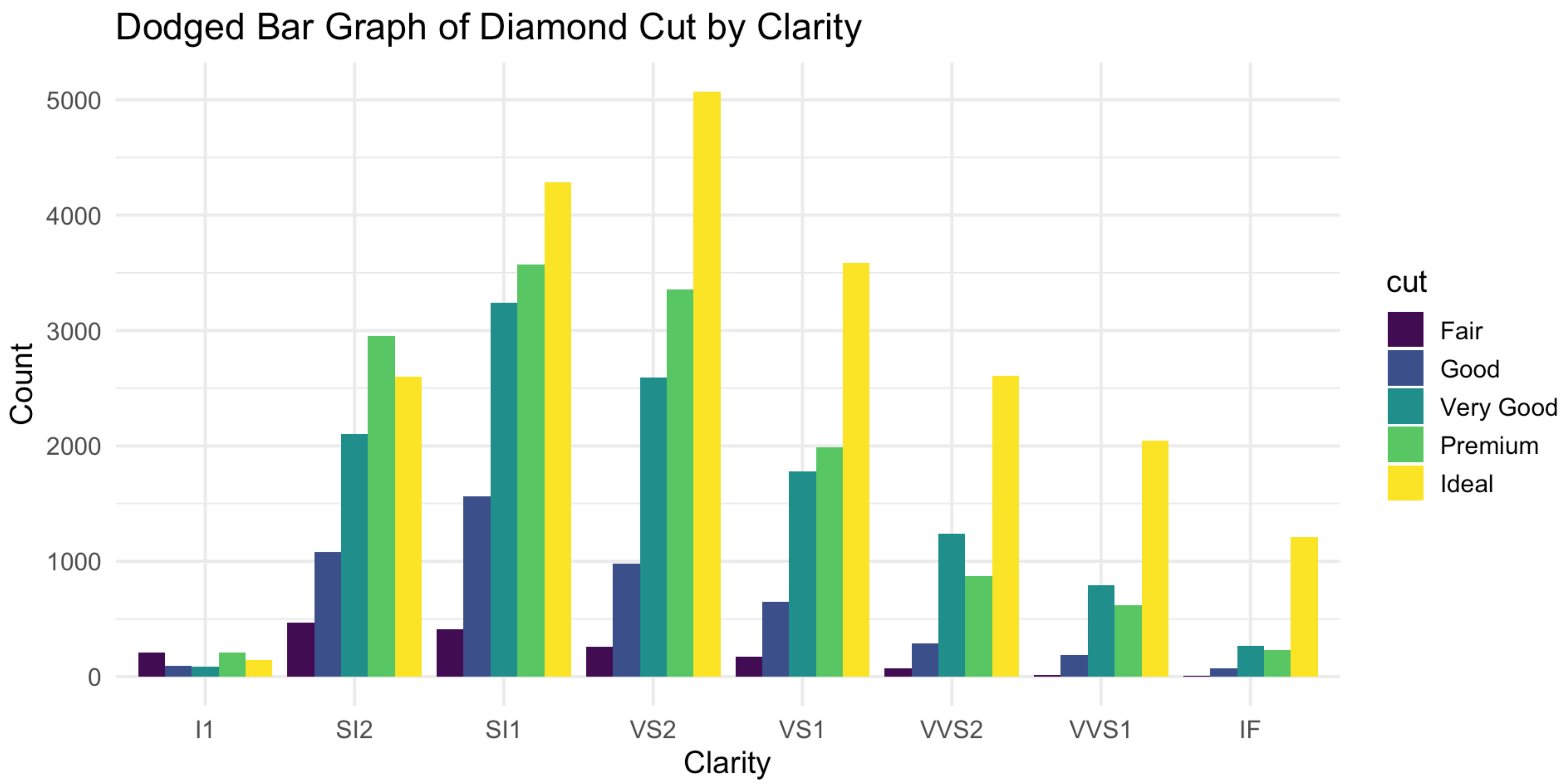
You can combine both color and shape to be more effective



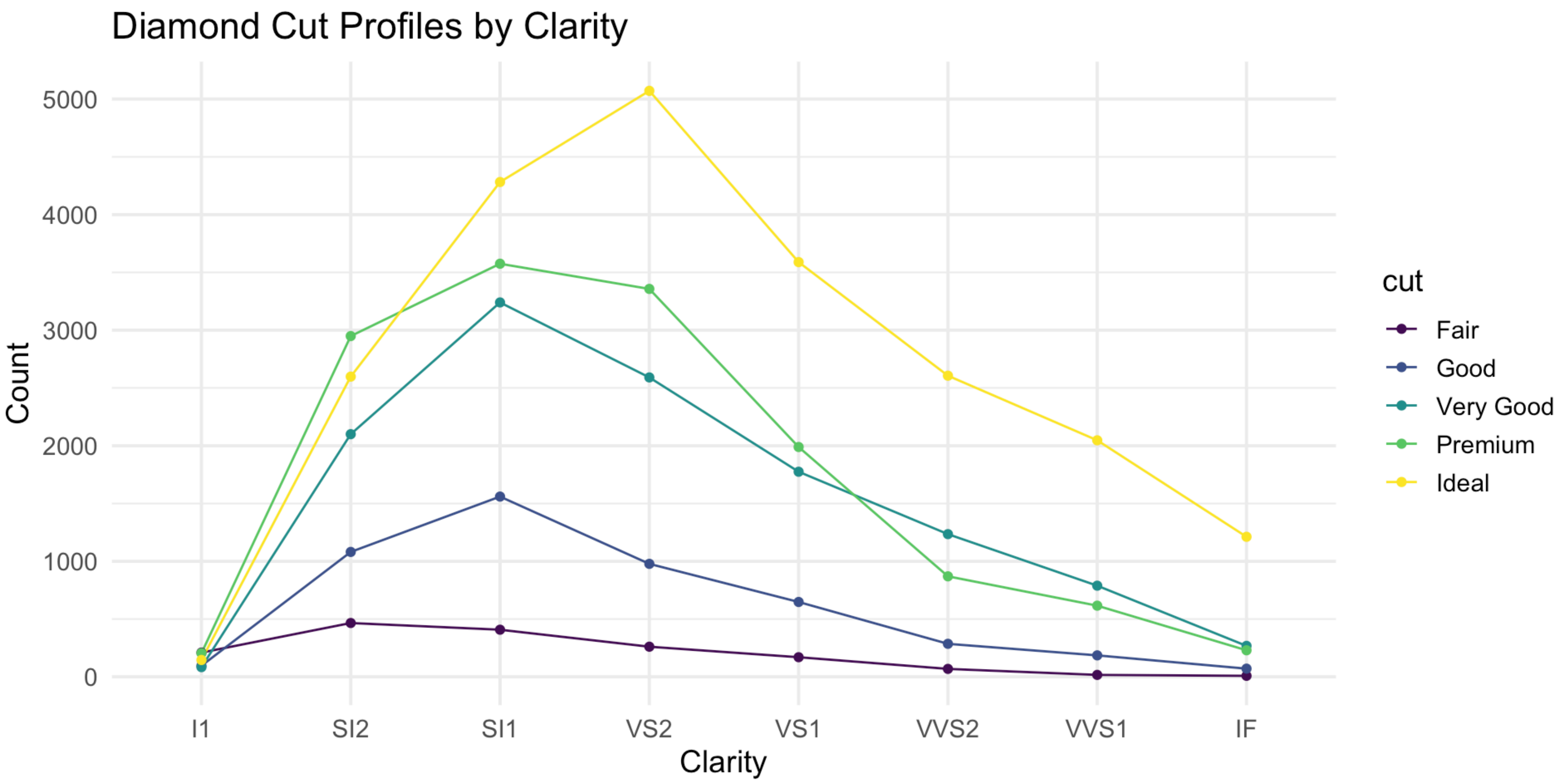
Use law of closure to group similar data



Law of proximity: we see elements near each other as part of the same object



Still worse than parallel coordinate plot



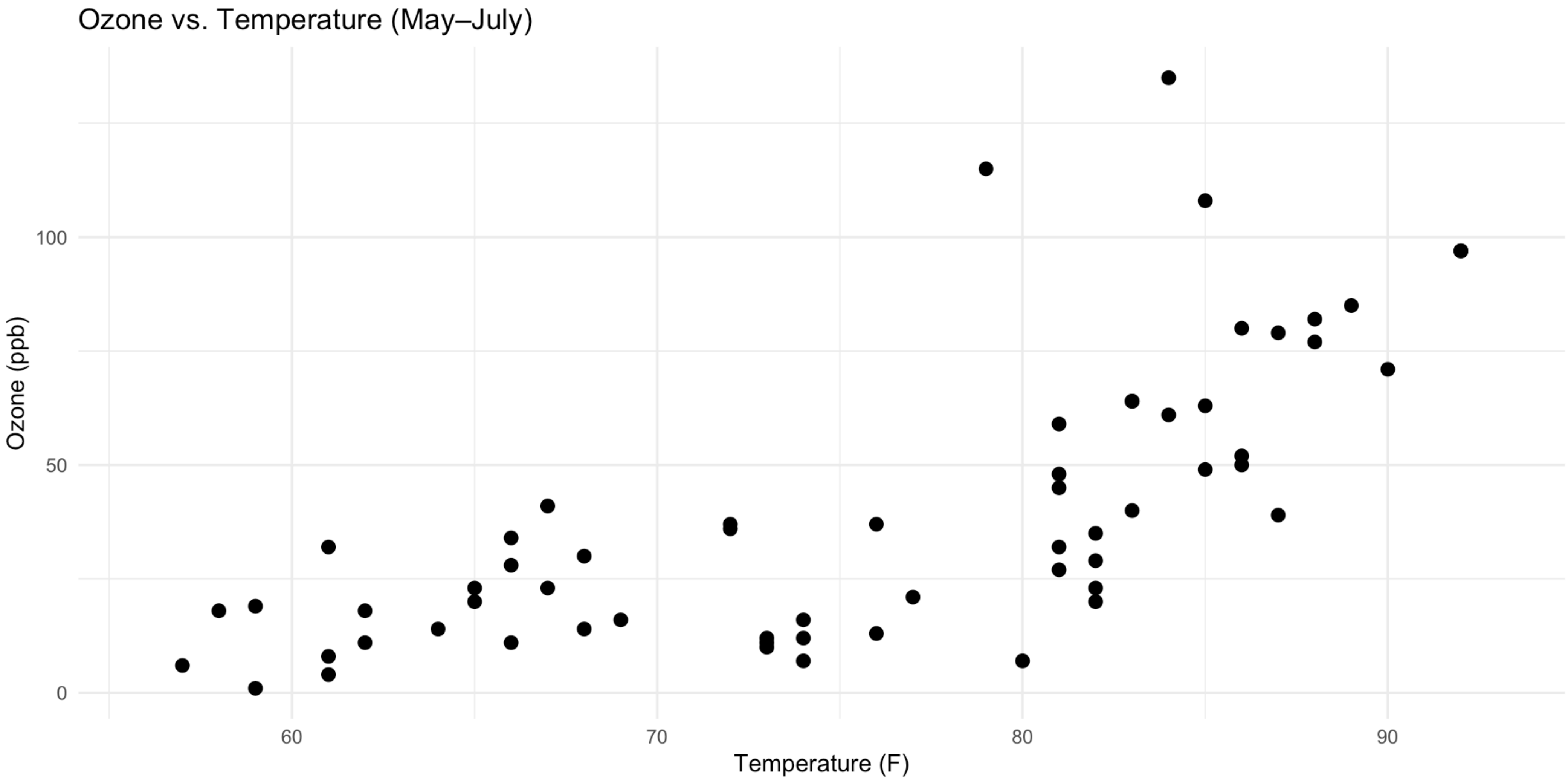
Cleveland’s three visual operations of pattern perception

🎯 **Detection**: *Recognizing that a geometric object encodes a physical value.*

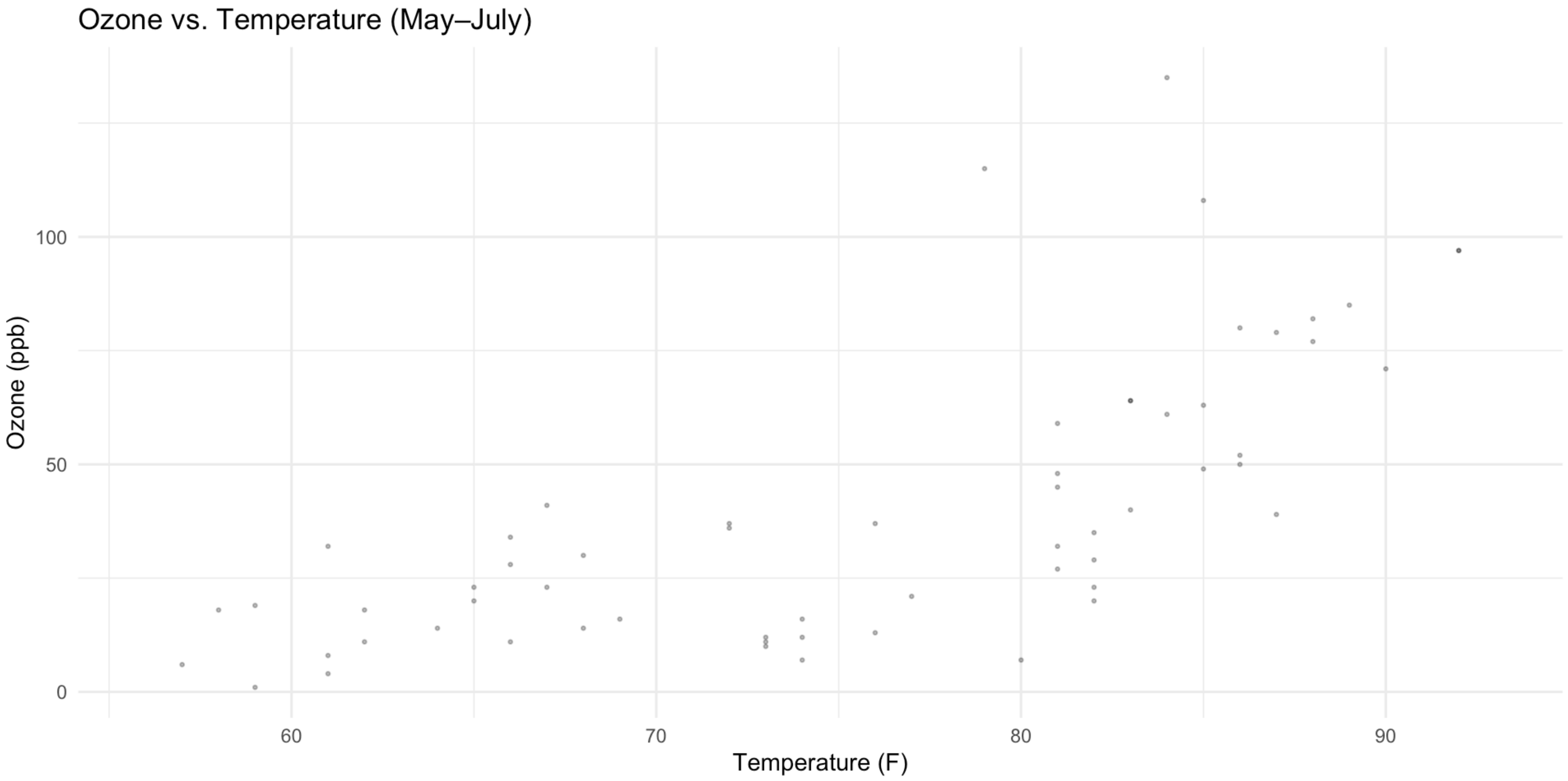
🧩 **Assembly**: *Grouping detected graphical elements into patterns.*

📏 **Estimation**: *Visually assessing the relative magnitude of two or more values.*

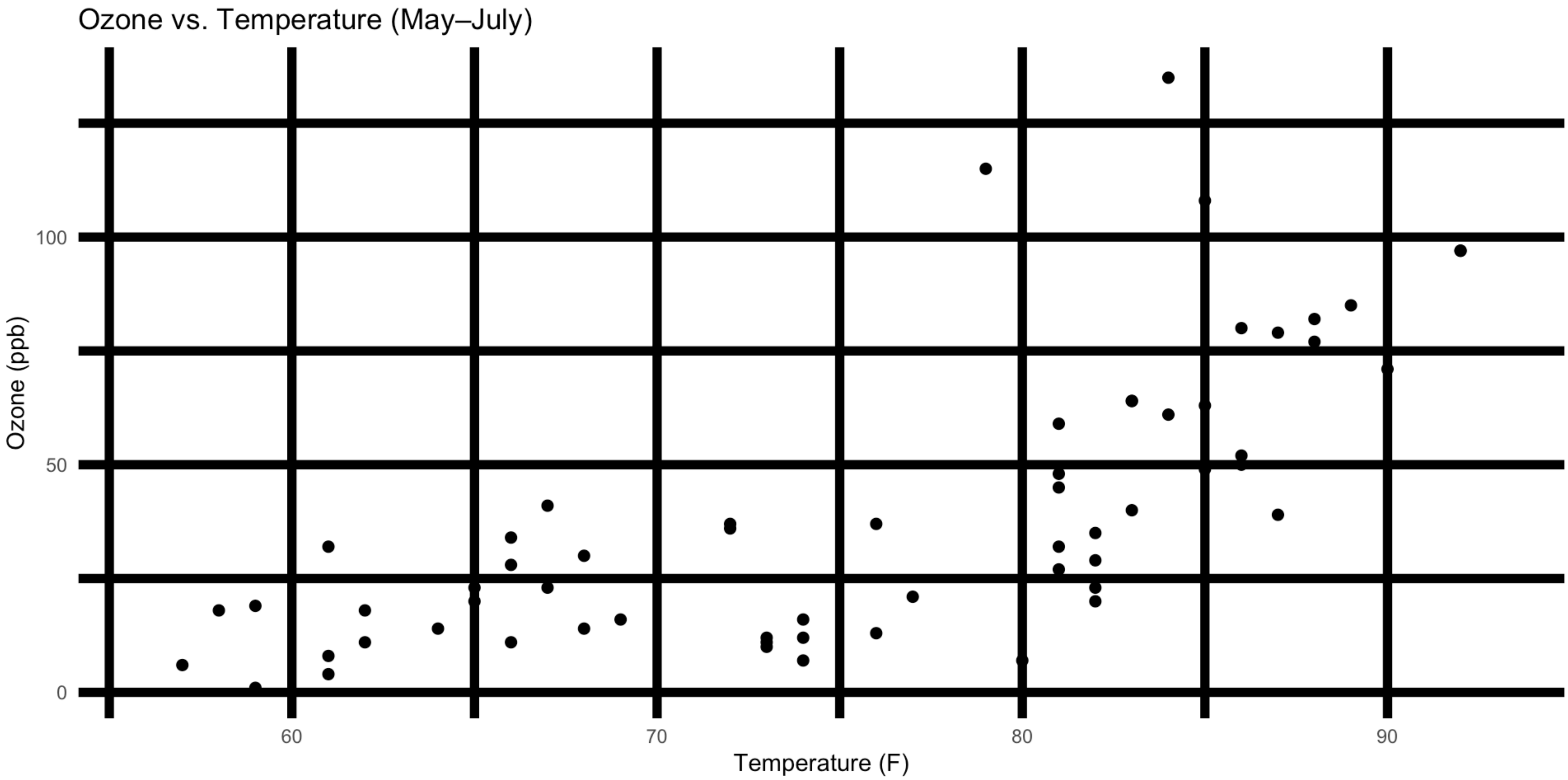
Detection should be trivial, don’t make it hard



Detection should be trivial, don’t make it hard



Detection should be trivial, don't make it hard



☕ Take a Break

~ This is the end of part 1 ~

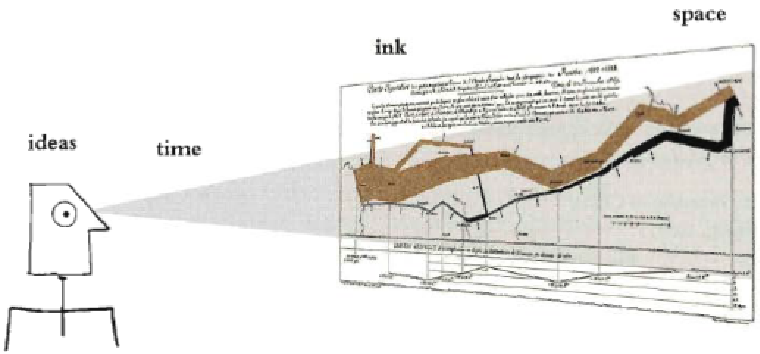
05:00

Outline for today

- How human see data
- Data-Ink Maximization and Graphical Redesign
- Design considerations for different types of intended audience

Principles of Graphical Excellence

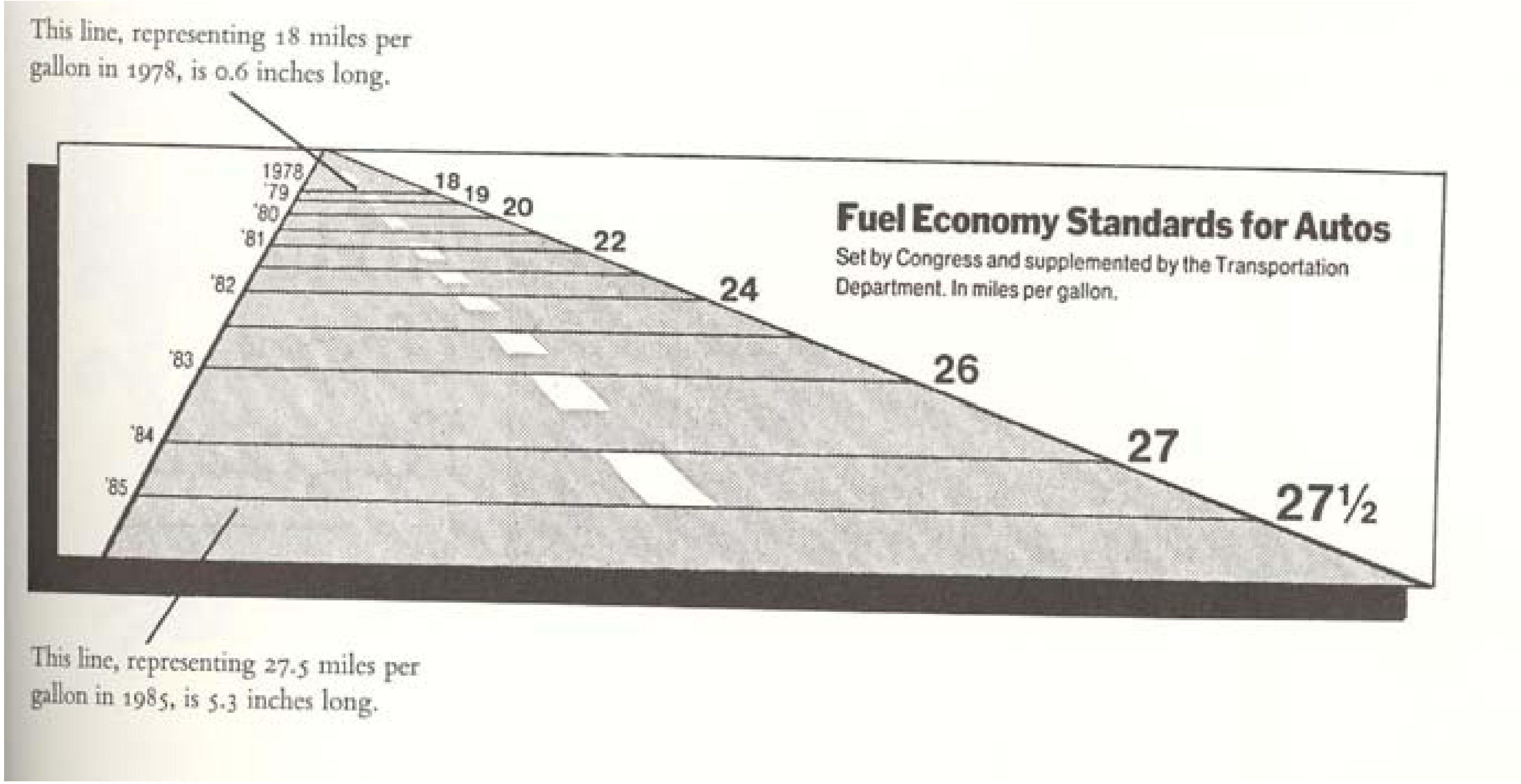
- Graphical excellence is the well-designed presentation of interesting data - a matter of *substance*, of *statistics*, and of *design*.
- Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency.
- Graphical excellence is that which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.



- Graphical excellence is nearly always multivariate.
- Graphical excellence requires telling the truth about the data.

Lie factor

$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

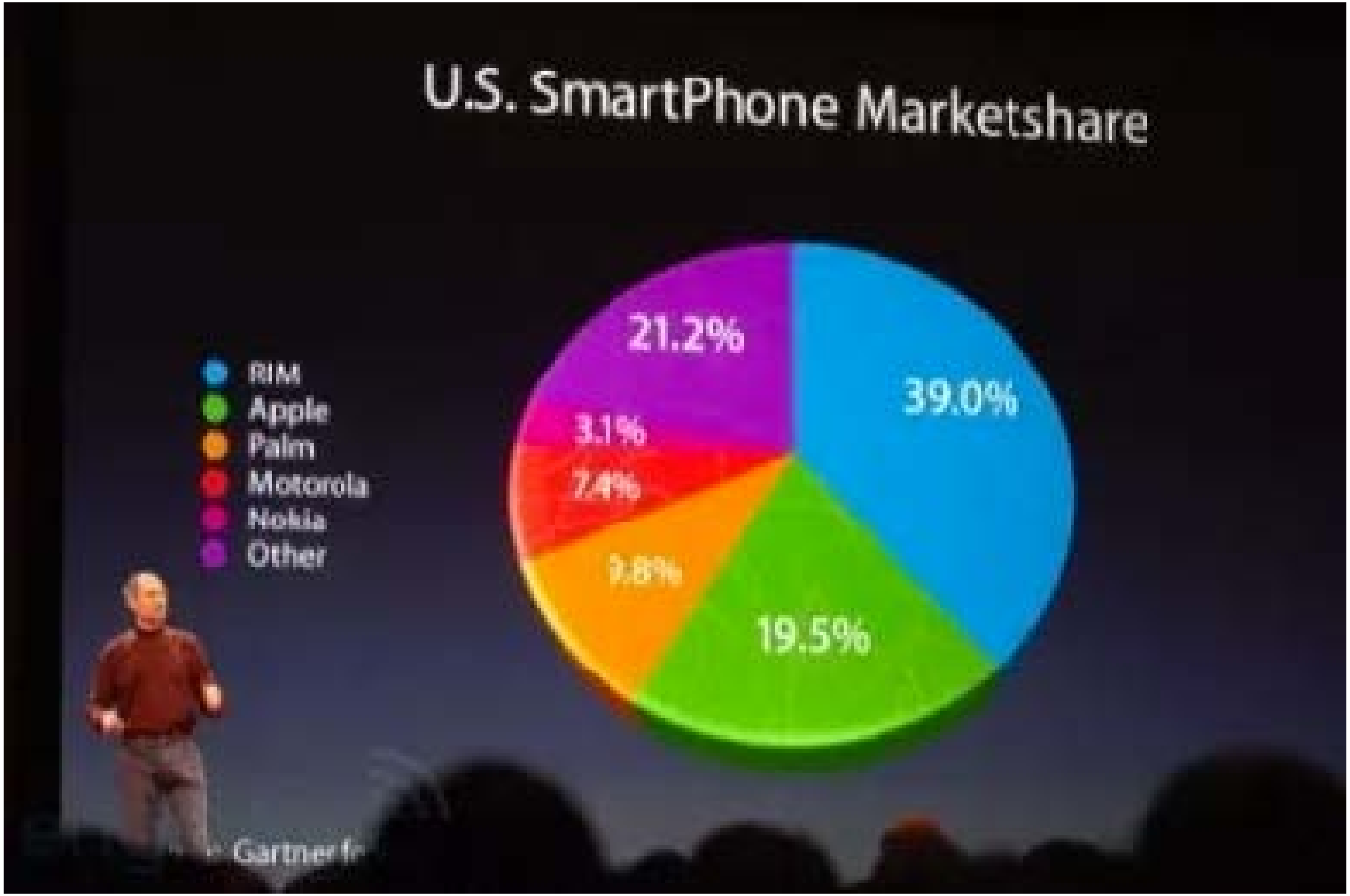


This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

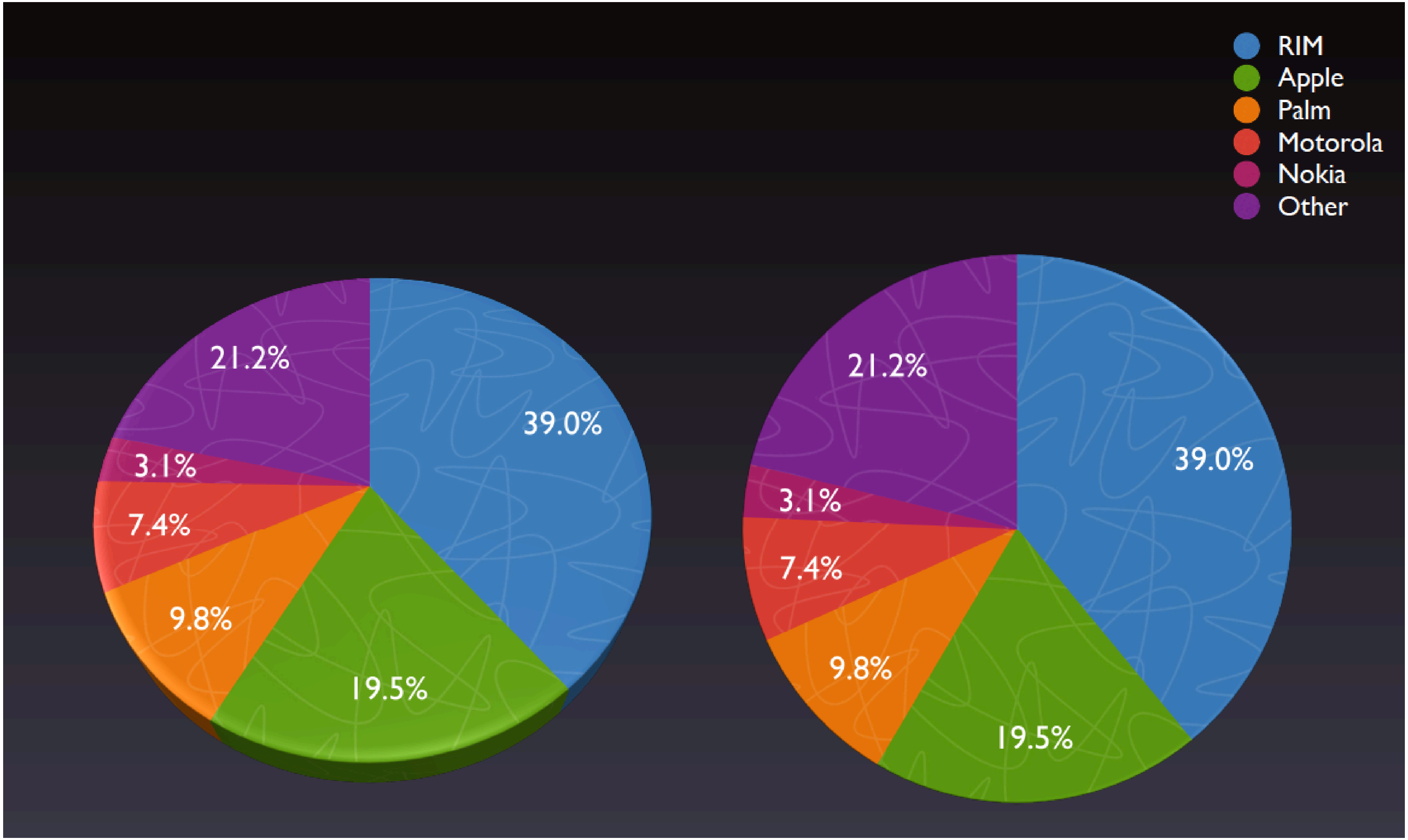
Can you calculate the lie factor in this graph?

Why are 3D graphs bad?

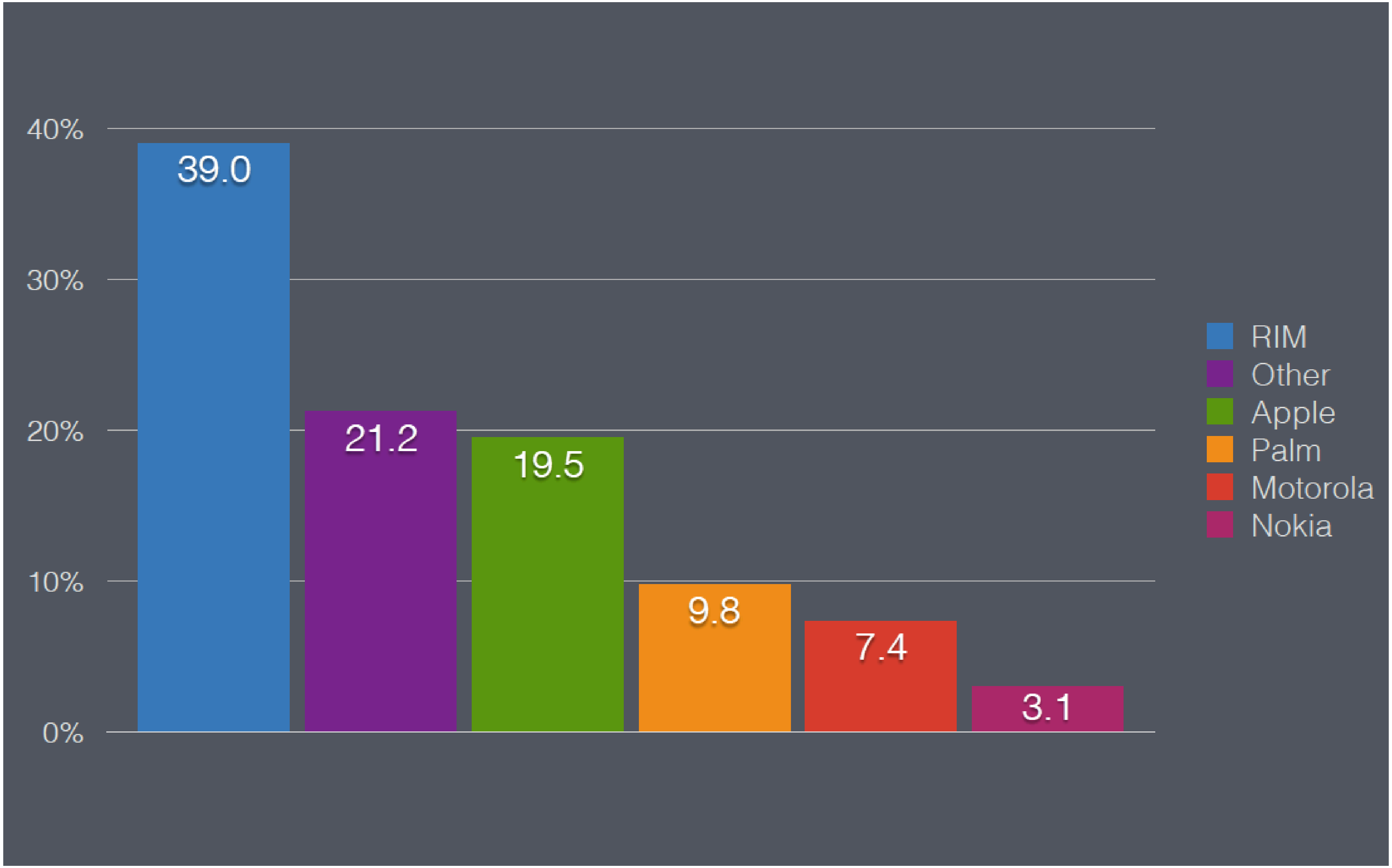


Source: the Guardian, 2008

How should the data be plotted?



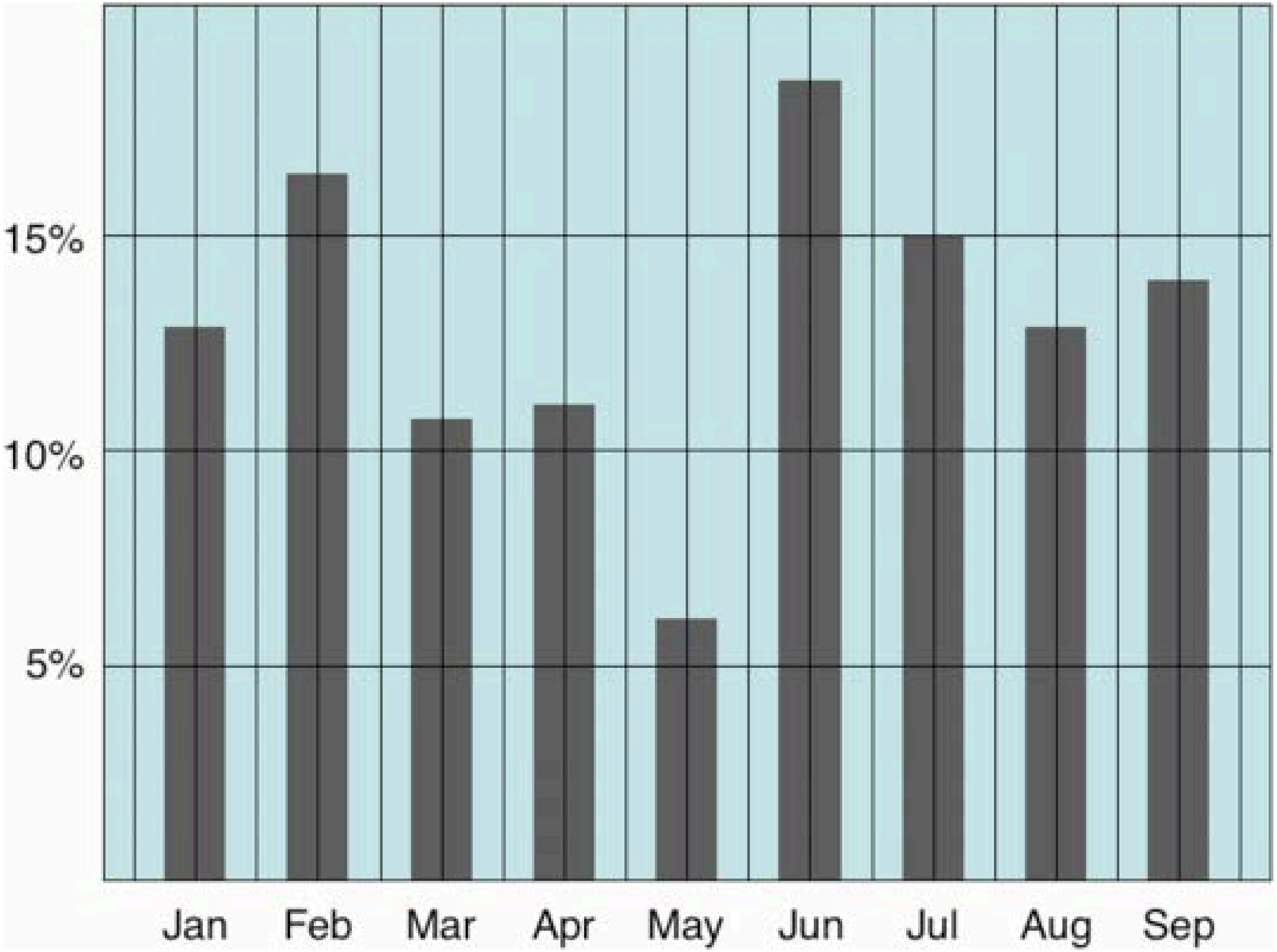
Or even better



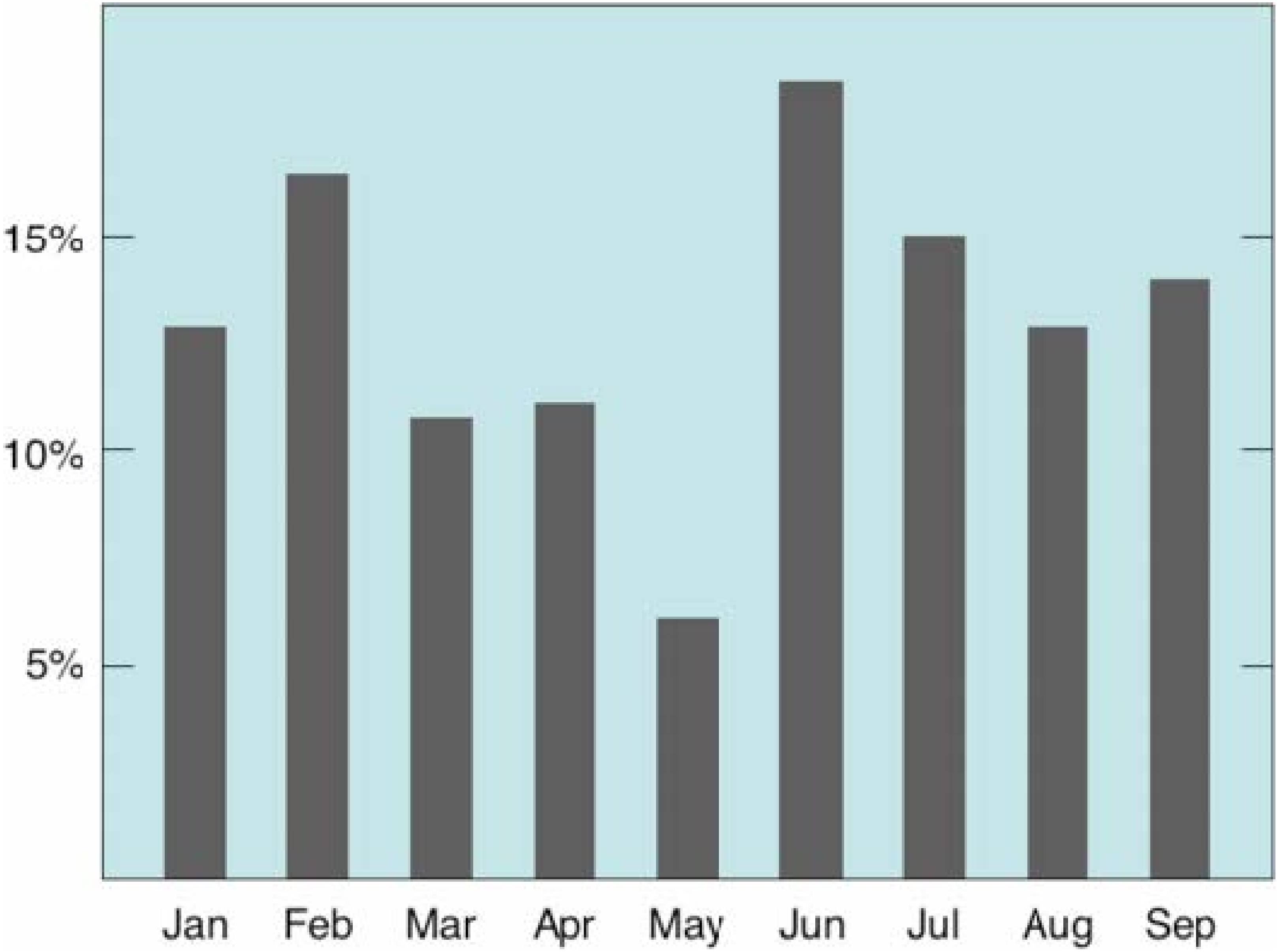
Maximize Data-Ink Ratio

$$\begin{aligned}\text{Data-Ink Ratio} &= \frac{\text{Data ink}}{\text{Total ink used in graphic}}\\ &= \text{proportion of a graphic's ink devoted to the}\\ &\ \text{non-redundant display of data-information}\\ &= 1 - \frac{\text{Redundant ink}}{\text{Total ink used in graphic}}\end{aligned}$$

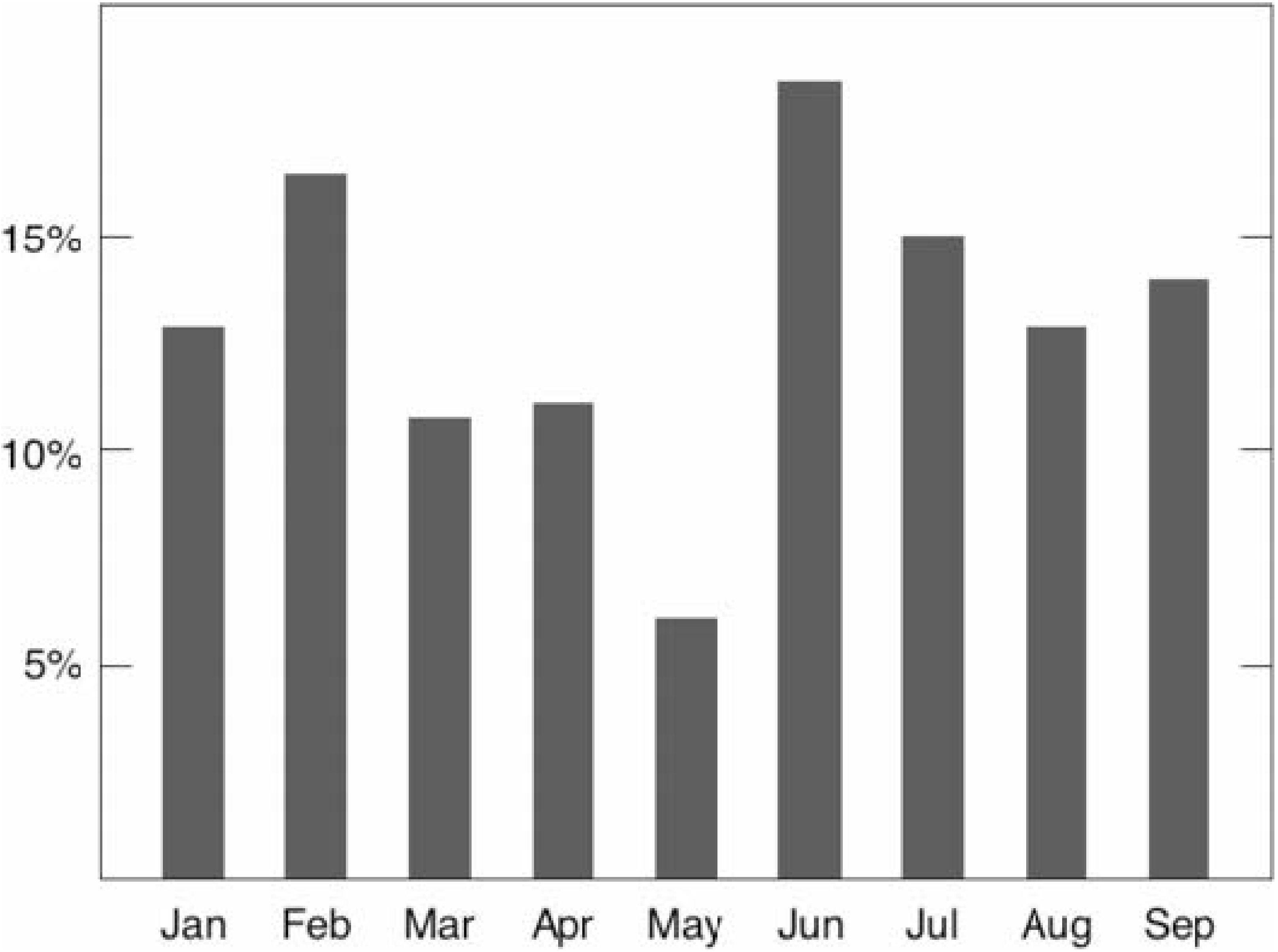
Avoid junk chart



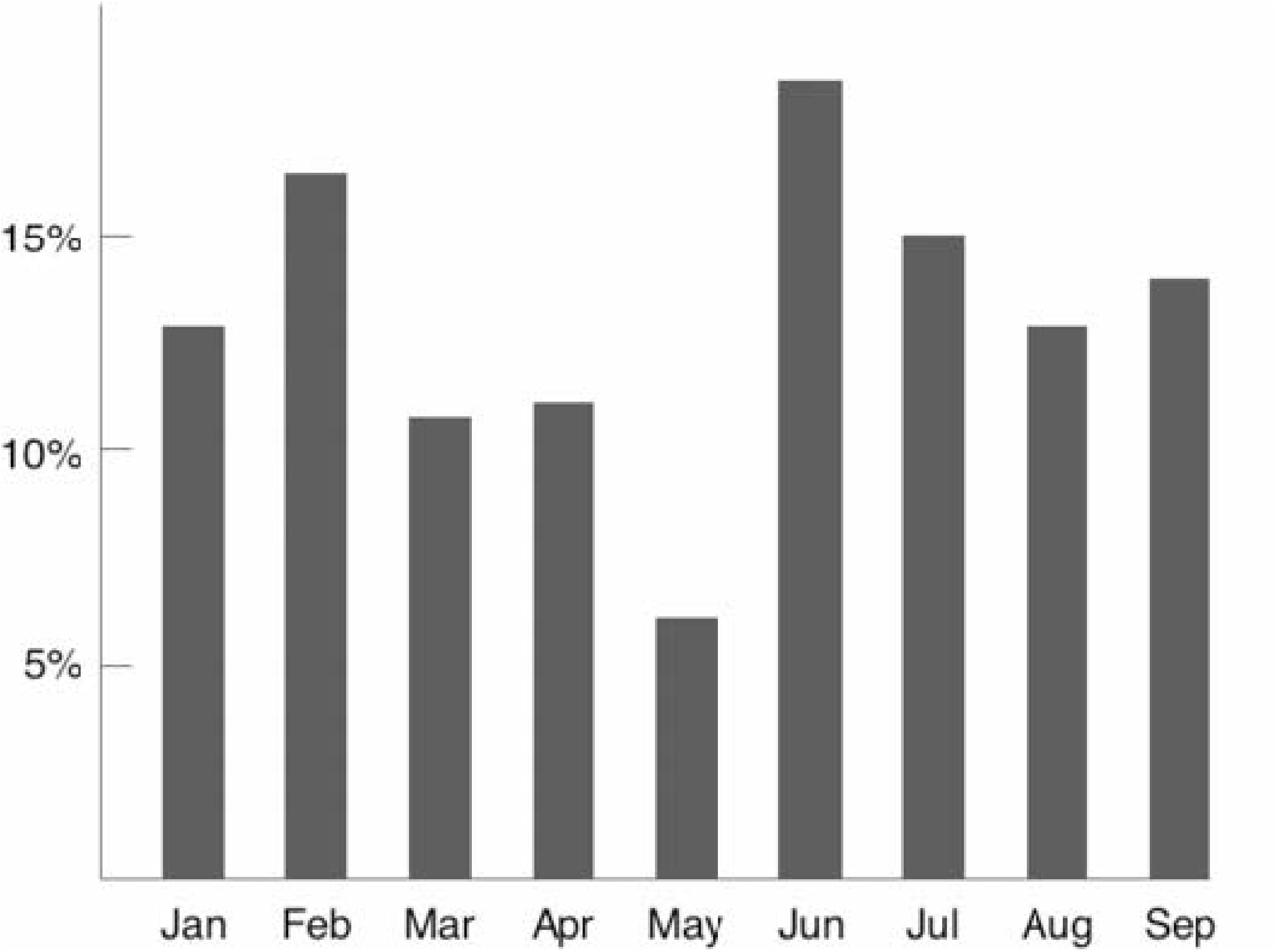
Avoid junk chart



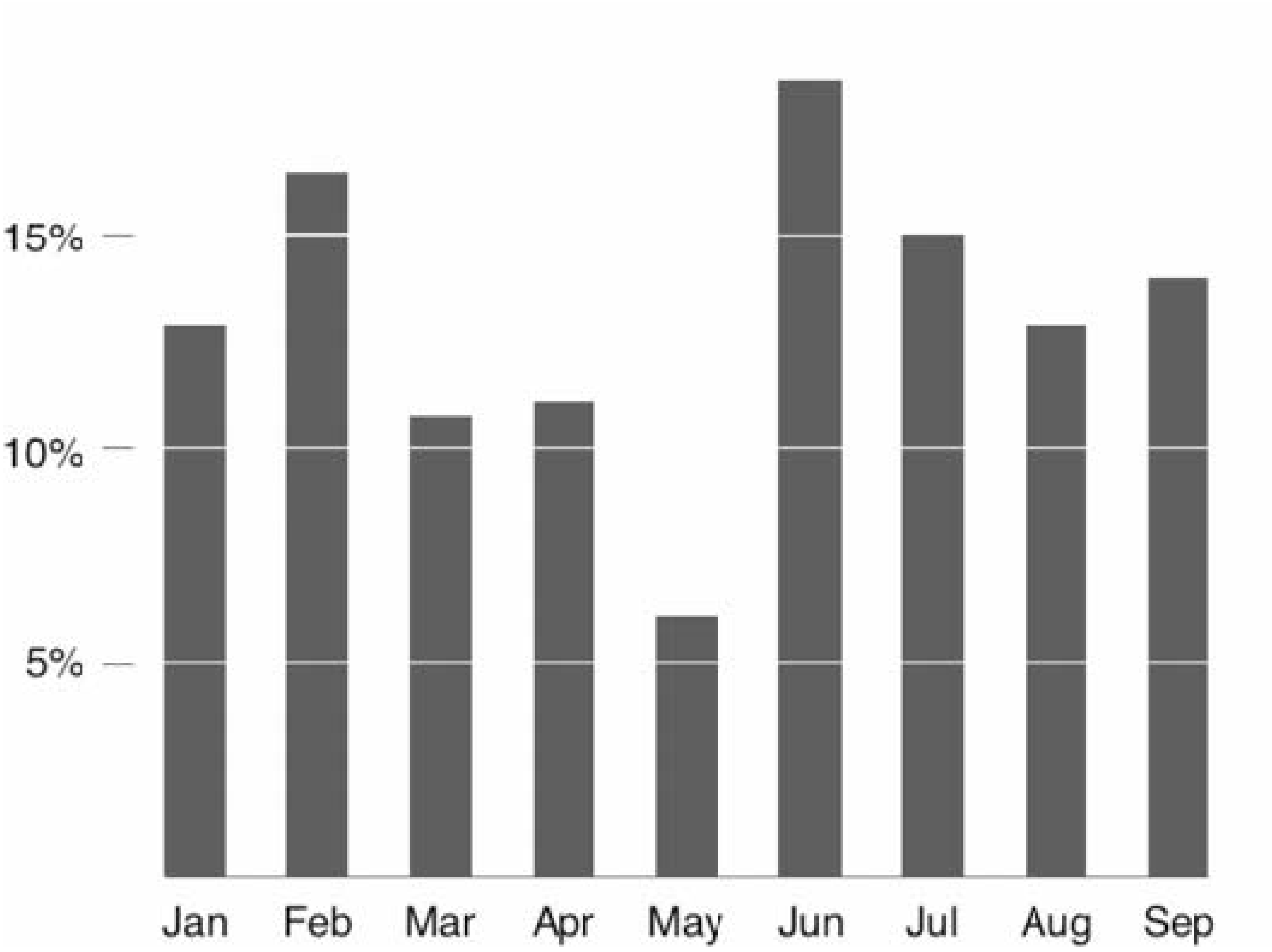
Avoid junk chart



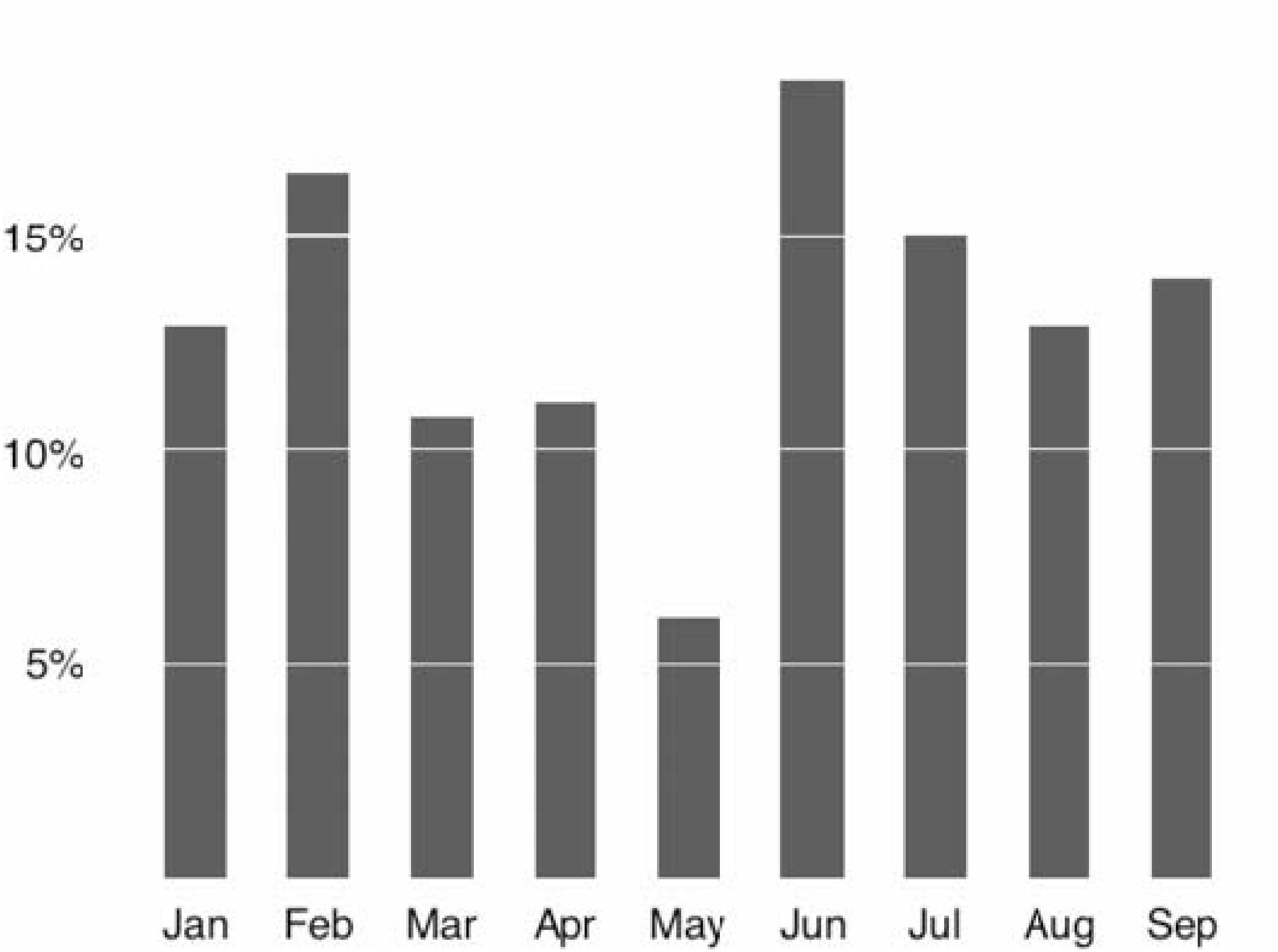
Avoid junk chart



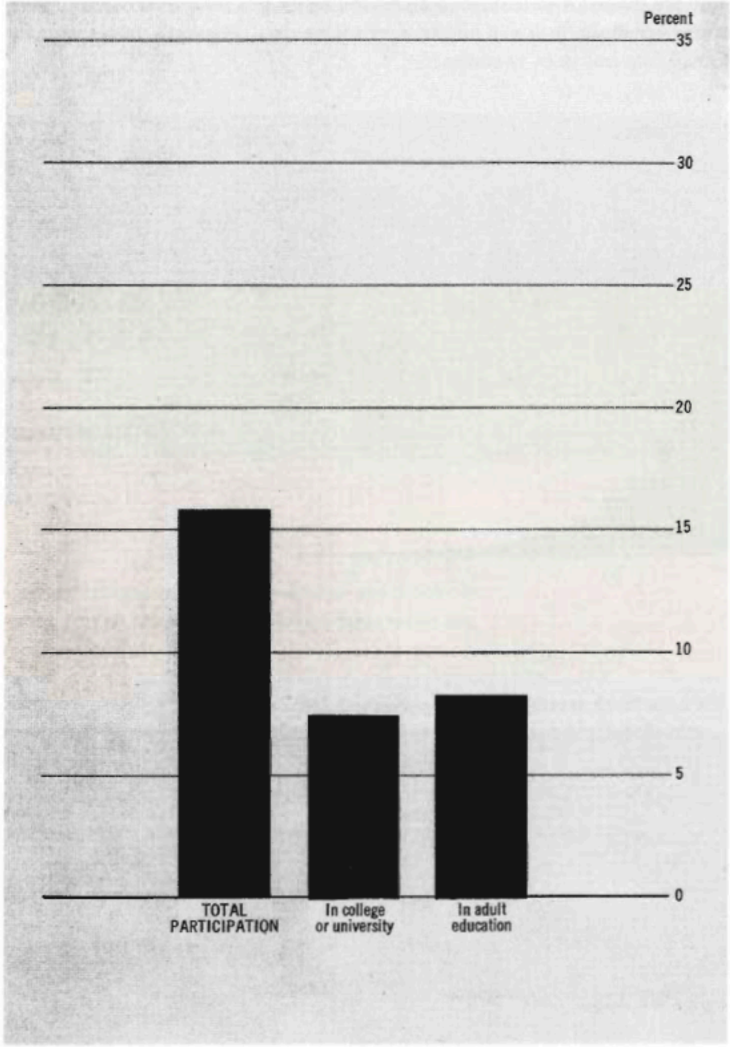
Avoid junk chart



Avoid junk chart



Data density in graphical practice



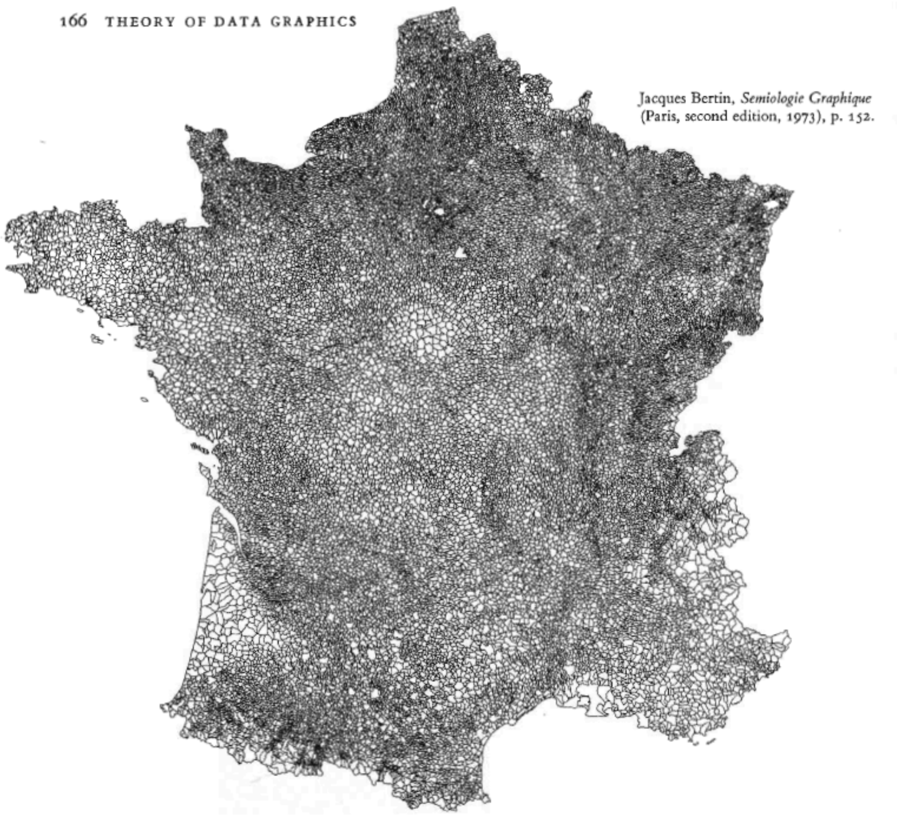
Office of Management and Budget

Social Indicators, 1973

$$\text{data density of a graphic} = \frac{\text{number of entries in data matrix}}{\text{area of data graphic}}$$

$$\text{data density} = \frac{\text{2 data points}}{\text{graph covres 26.5 square inch}}$$
$$= 0.15 \text{ numbers per square inch}$$

Data density in graphical practice



166 THEORY OF DATA GRAPHICS

Jacques Bertin, *Semiologie Graphique* (Paris, second edition, 1973), p. 152.

Jacques Bertin, *Semiologie Graphique*, 1973

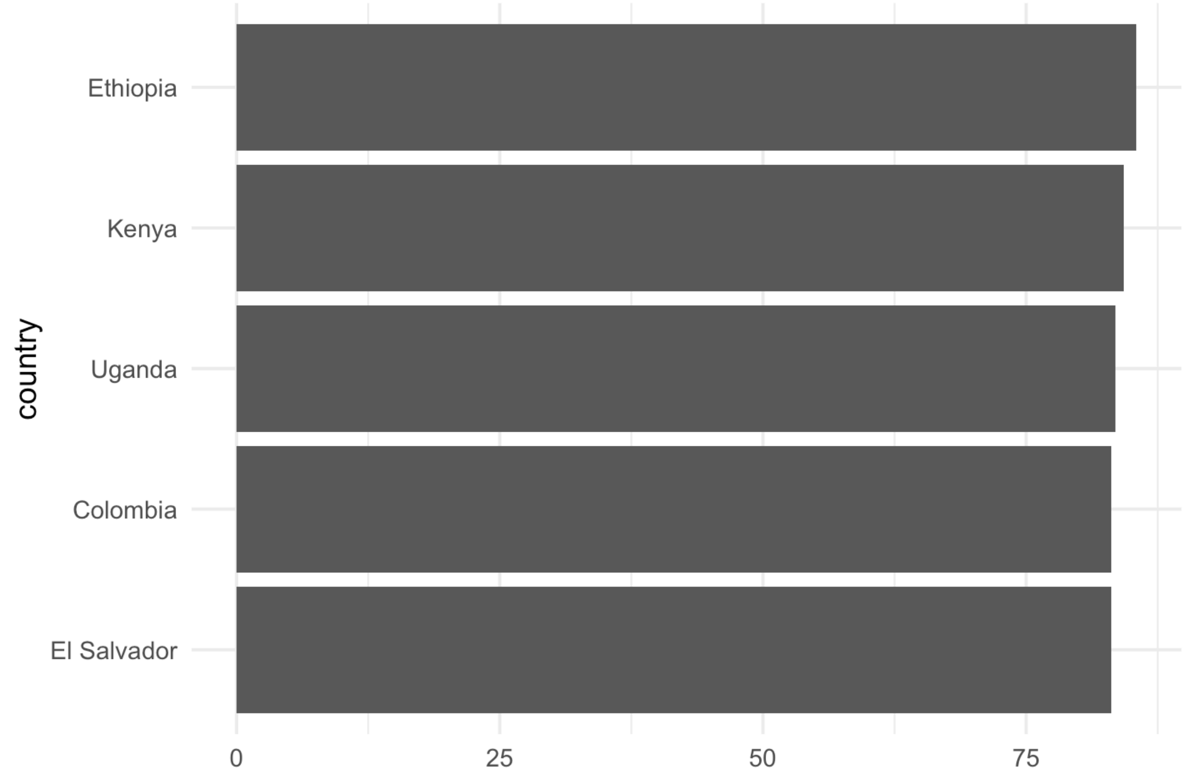
$$\text{data density of a graphic} = \frac{\text{number of entries in data}}{\text{area of data graphic}}$$

$$\text{data density} = \frac{\text{240,000 data points}}{\text{graph covres 27 square inch}}$$
$$= 9,000 \text{ numbers per square inch}$$

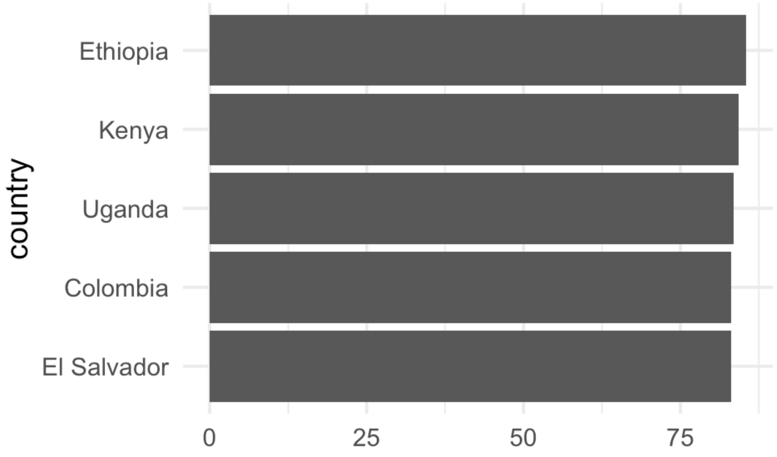
How to create high-information graphics design?

Graphics can be shrunk way down

Default size



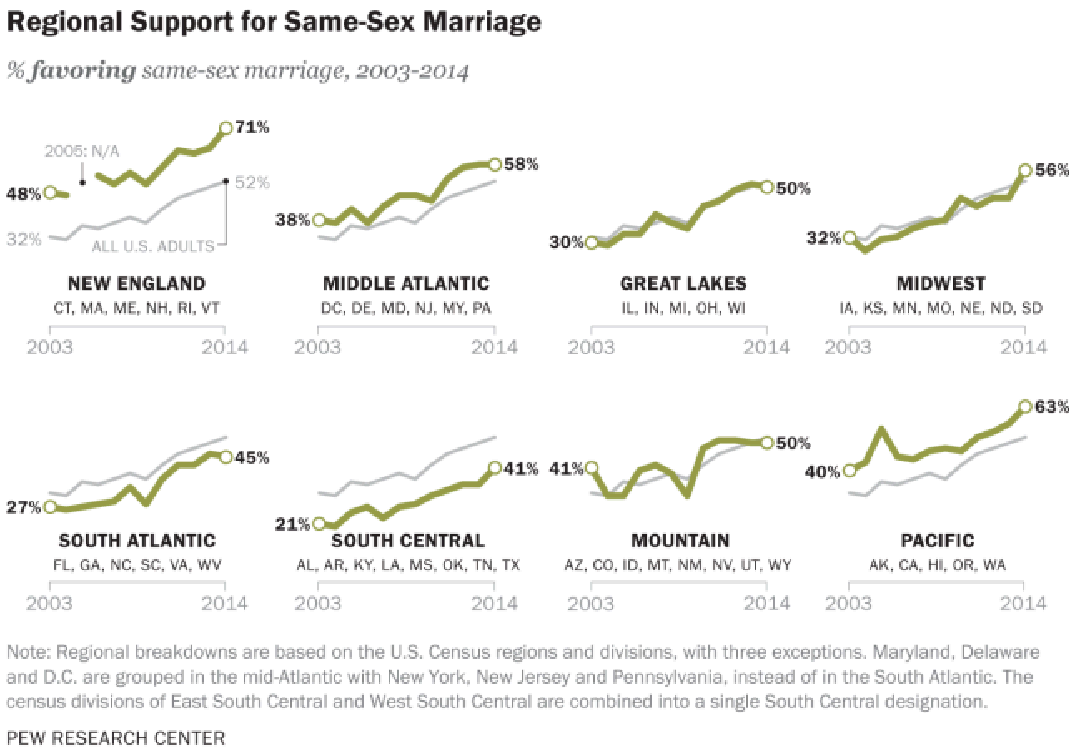
Appropriate size



Small Multiples

> “Small multiples resemble the frames of a movie: a series of graphics, showing the same combination of variables, indexed by changes in another variable.”

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.



Regional Support for Same-Sex Marriage

% favoring same-sex marriage, 2003-2014

Note: Regional breakdowns are based on the U.S. Census regions and divisions, with three exceptions. Maryland, Delaware and D.C. are grouped in the mid-Atlantic with New York, New Jersey and Pennsylvania, instead of in the South Atlantic. The census divisions of East South Central and West South Central are combined into a single South Central designation.

PEW RESEARCH CENTER

Pew Research Center

Well-designed small multiples are

- inevitably comparative
- deftly multivariate
- shrunken, high-density graphics
- usually based on a large data matrix
- draw almost entirely with data-ink
- efficient in interpretation
- often narrative in content, showing shifts in the relationship between variables as the index variable changes (thereby revealing interaction or multiplicative effects)

Outline for today

- How human see data
- Data-Ink Maximization and Graphical Redesign
- **Design considerations for different types of intended audience**

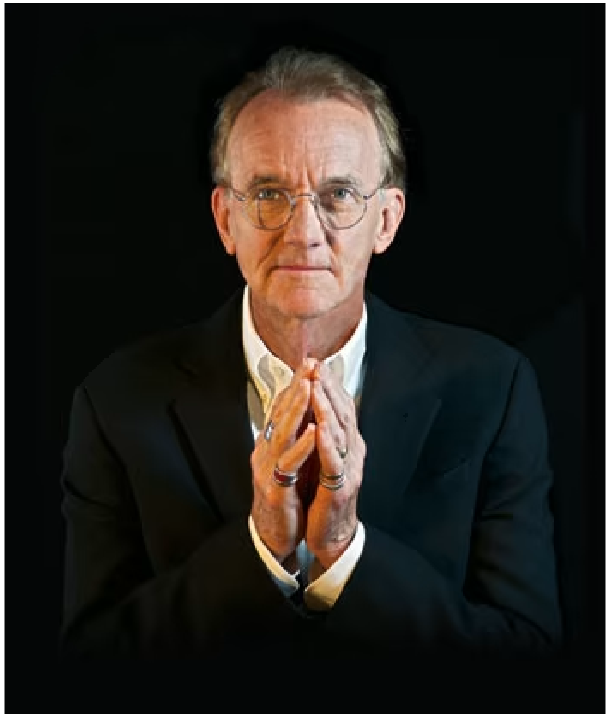
Audience dimensions

Audience may vary by:

- **Domain knowledge**: the field of study
- **Statistical literacy**: the level of knowledge
- **Time constraints**: the time available to read the data
- **Cognitive load**: the ability to process large amount of information
- **Expectations for interactivity or aesthetics**

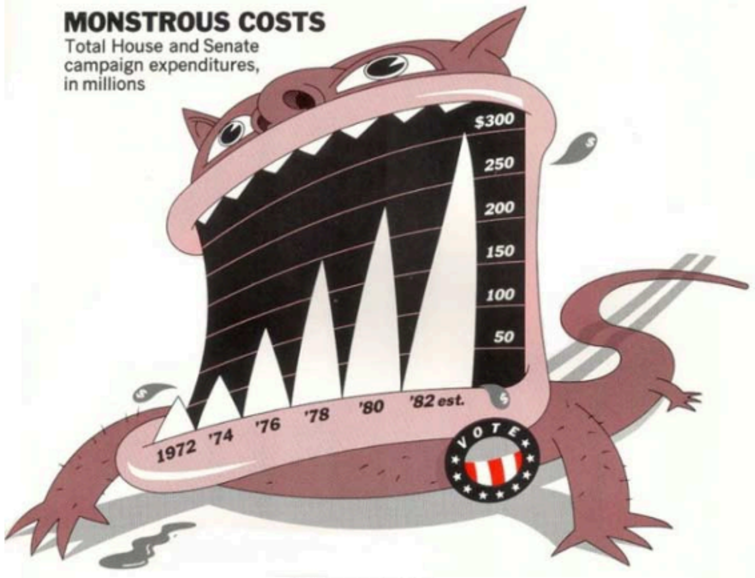
Tufte's design principles

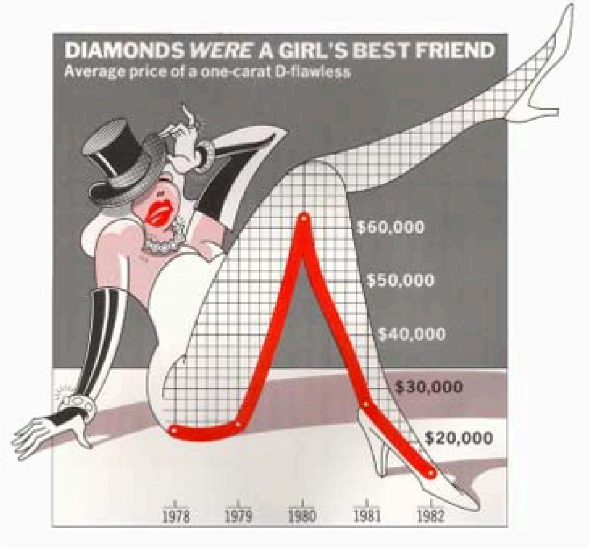
- Graphical integrity
- The Lie Factor
- Maximize data-ink ratio
- Avoid chart junk

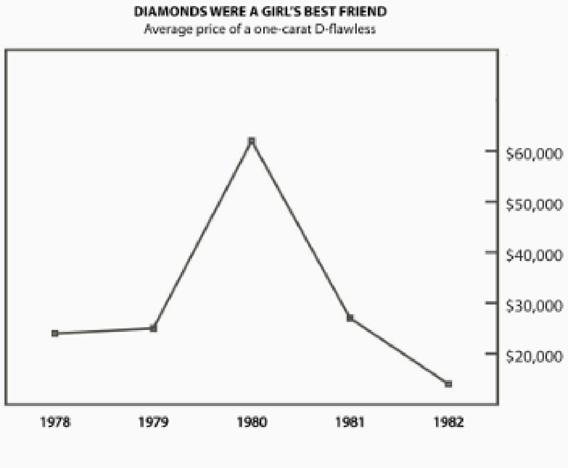


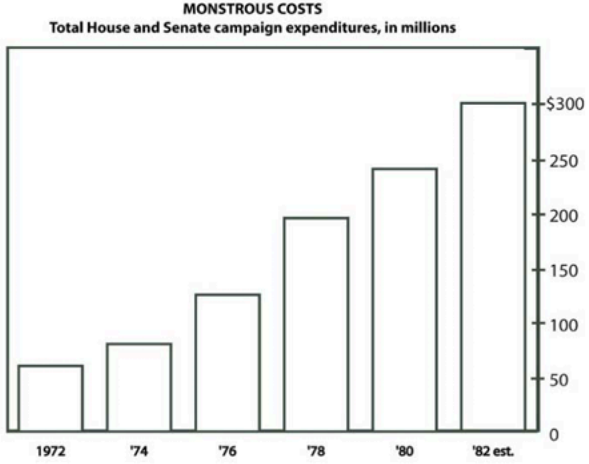
Most useful for analytical or technical audience, e.g. scientists, engineers, and data analysts. Less useful for the general public or media campaigns.

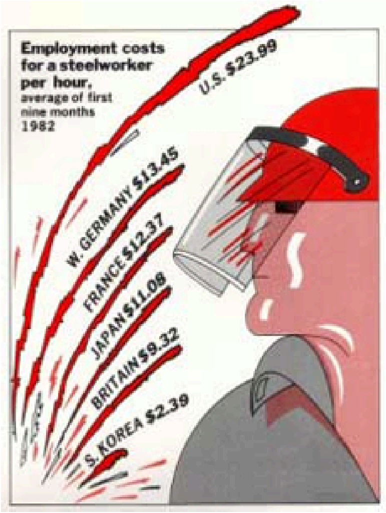
Useful junk

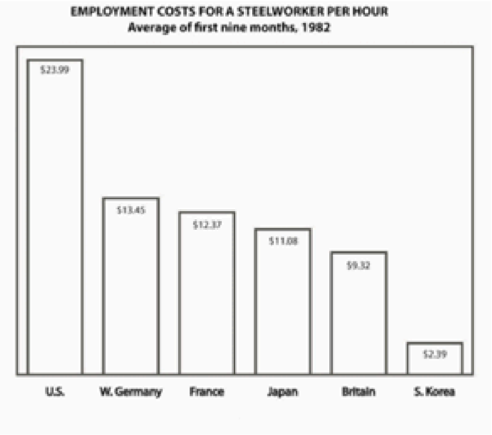




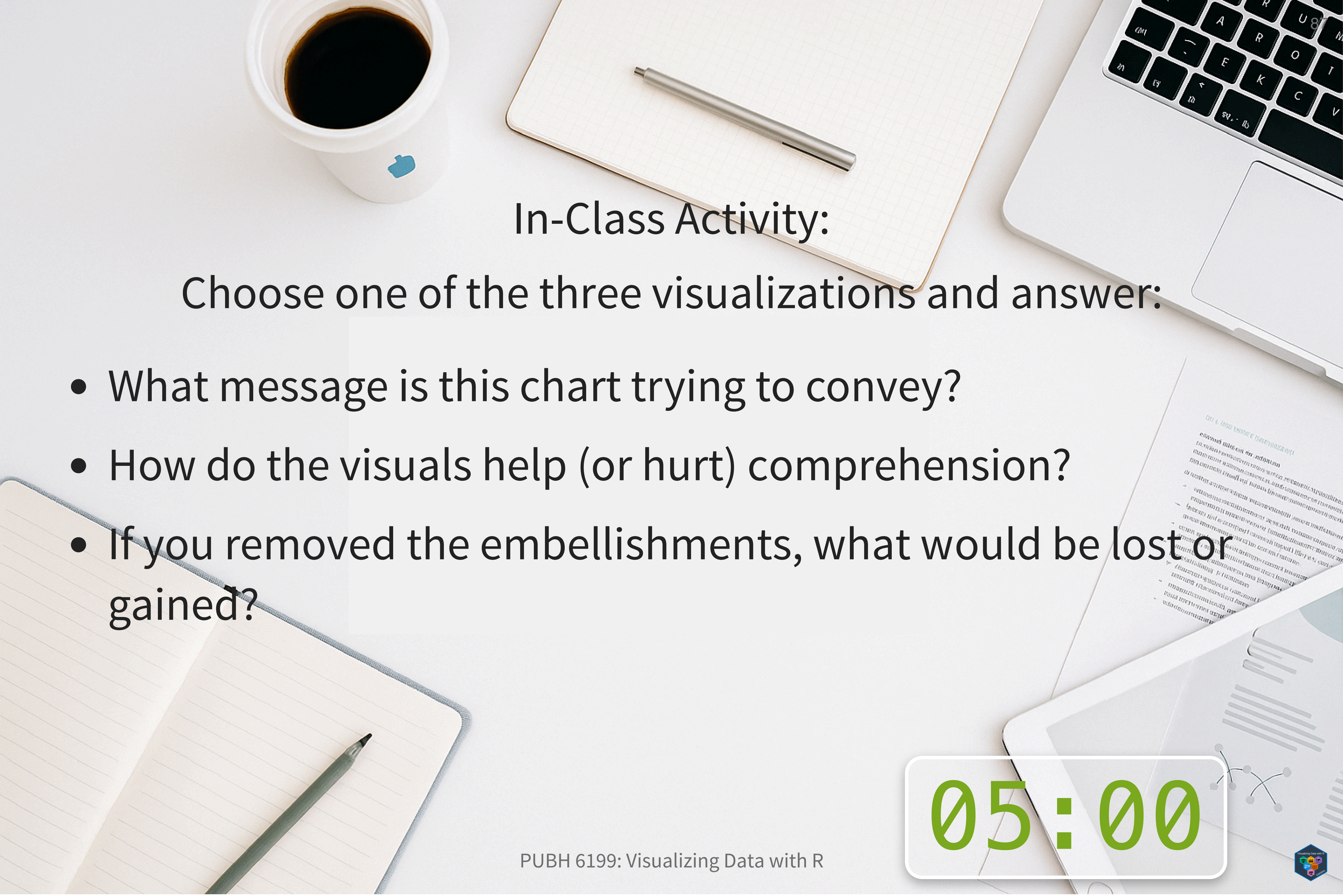








Bateman et al, CHI 2010: Graphs



In-Class Activity:

Choose one of the three visualizations and answer:

- What message is this chart trying to convey?
- How do the visuals help (or hurt) comprehension?
- If you removed the embellishments, what would be lost or gained?

05:00

Data accessibility for individuals with intellectural or developmental disabilities

Data accessibility for individuals with color blindedness

Color blindness affects approximately 1 in 12 men and 1 in 200 women. To ensure your visualizations remain accessible:

- **Avoid red-green or red-brown combinations**
- **Use colorblind-friendly palettes**, such as `viridis`, `Okabe-Ito`, or `Color Universal Design (CUD)`
- **Add texture, shape, or direct labels** to differentiate groups beyond color
- **Test your charts** with tools like `colorblindr`
- **Use contrast checkers** to ensure sufficient visual separation

> Designing with color blindness in mind improves clarity for everyone.

End-of-Class Survey

✎ Fill out the end-of-class survey

~ This is the end of Lecture 2 ~

10:00