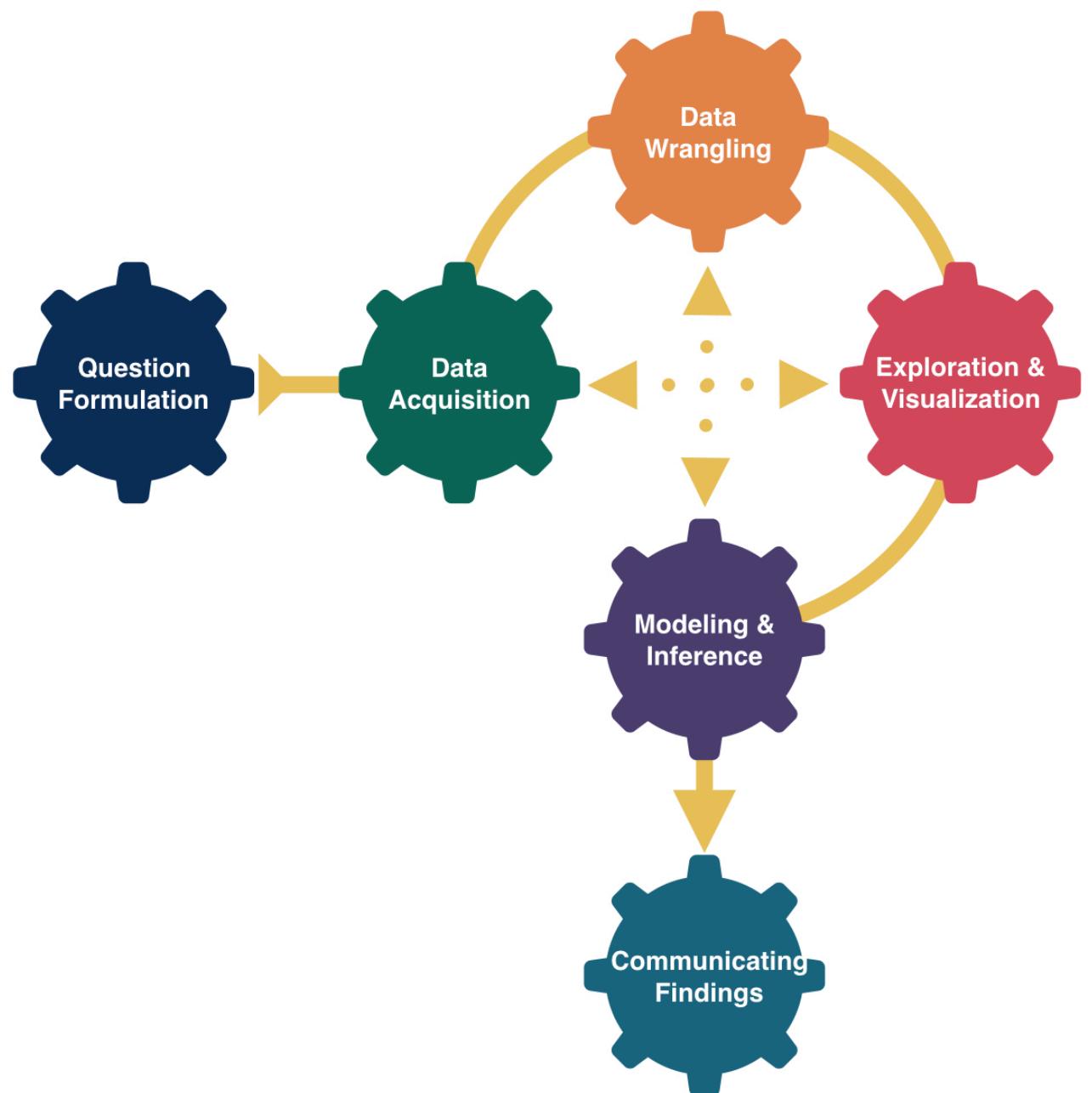


Data Visualization

Kelly McConville

DATA 100

Week 2 | Spring 2026



Announcements

- Class in full swing:
 - P-set 1 due on Gradescope by noon on Thursday.
 - Visit office hours and then fill out [this form](#)
 - Panel on careers in data science and public health on Thursday at noon in Taylor 210.

Announcements

Careers at the Intersection: Health, Data & Life Science



Michael Castro '97



Carolyn Ritterson Lew '04



Bozena Morawski '04

Register on Handshake ➔



Thursday, January 29

noon - 1:00 p.m. + + + Taylor Hall 210

Join us for a conversation with alumni working in roles that blend healthcare, data analytics and life sciences. Attendees will leave with a better understanding of the diverse career paths available and the competencies needed to thrive in this rapidly evolving intersection. Lunch will be provided for attendees.

Bucknell
UNIVERSITY | Center for
Career Advancement

Bucknell
UNIVERSITY | Dominguez Center
for Data Science



Goals for Today

First Segment:

- Motivate data visualizations.
- Develop **language** to talk about the components of a graphic.
- Practice deconstructing graphics.
- Discuss good graphical practices.

Second Segment:

- Learn the general structure of **ggplot2**.
- Learn a few standard graphs for numerical/quantitative data:
 - **Histogram**: one numerical variable
 - **Side-by-side boxplot**: one numerical variable and one categorical variable
 - **Side-by-side violin plot**: one numerical variable and one categorical variable

Why construct a graph?

To **explore** the data.

To **summarize** the data.

To showcase **trends** and make **comparisons**.

To tell a compelling **story**.

Challenger

- On January 27th, 1986, engineers from Morton Thiokol recommended NASA delay launch of space shuttle *Challenger* due to cold weather.
 - Believed cold weather impacted the o-rings that held the rockets together.
 - Used 13 charts in their argument.
- After a two hour conference call, the engineer's recommendation was overruled due to lack of persuasive evidence and the launch proceeded.
- The Challenger exploded 73 seconds into launch.

Challenger

Here's one of those charts.

Blow By HISTORY

SRM-15 WORST

Blow-By

- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 Blow-By

- 2 CASE JOINTS ($30-40^\circ$)

SRM-13A, 15, 16A, 18, 23A 24A

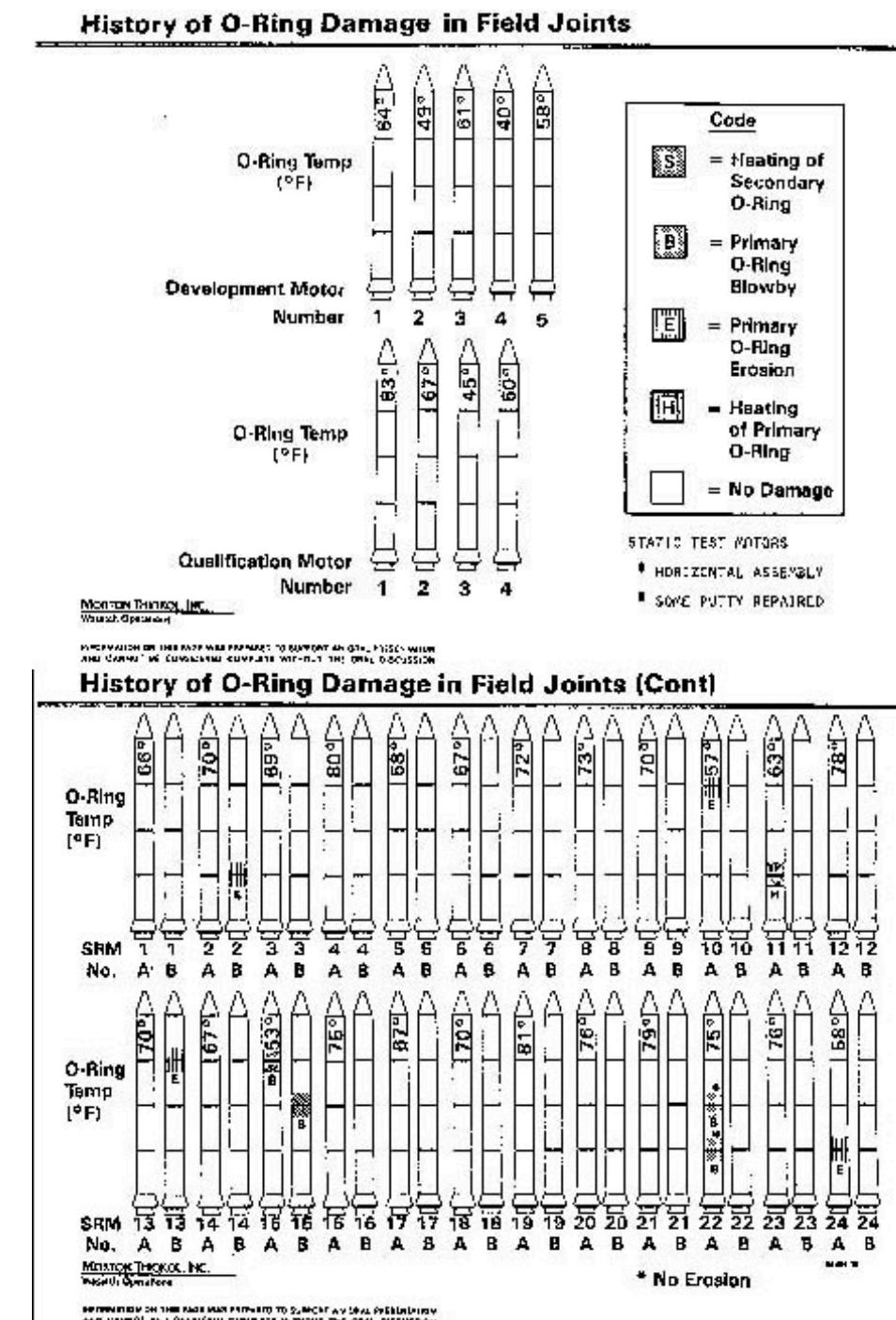
- NOZZLE Blow-By

HISTORY OF O-RING TEMPERATURES (DEGREES - F)

MOTOR	MBT	AMB	O-RING	WIND
DM-4	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

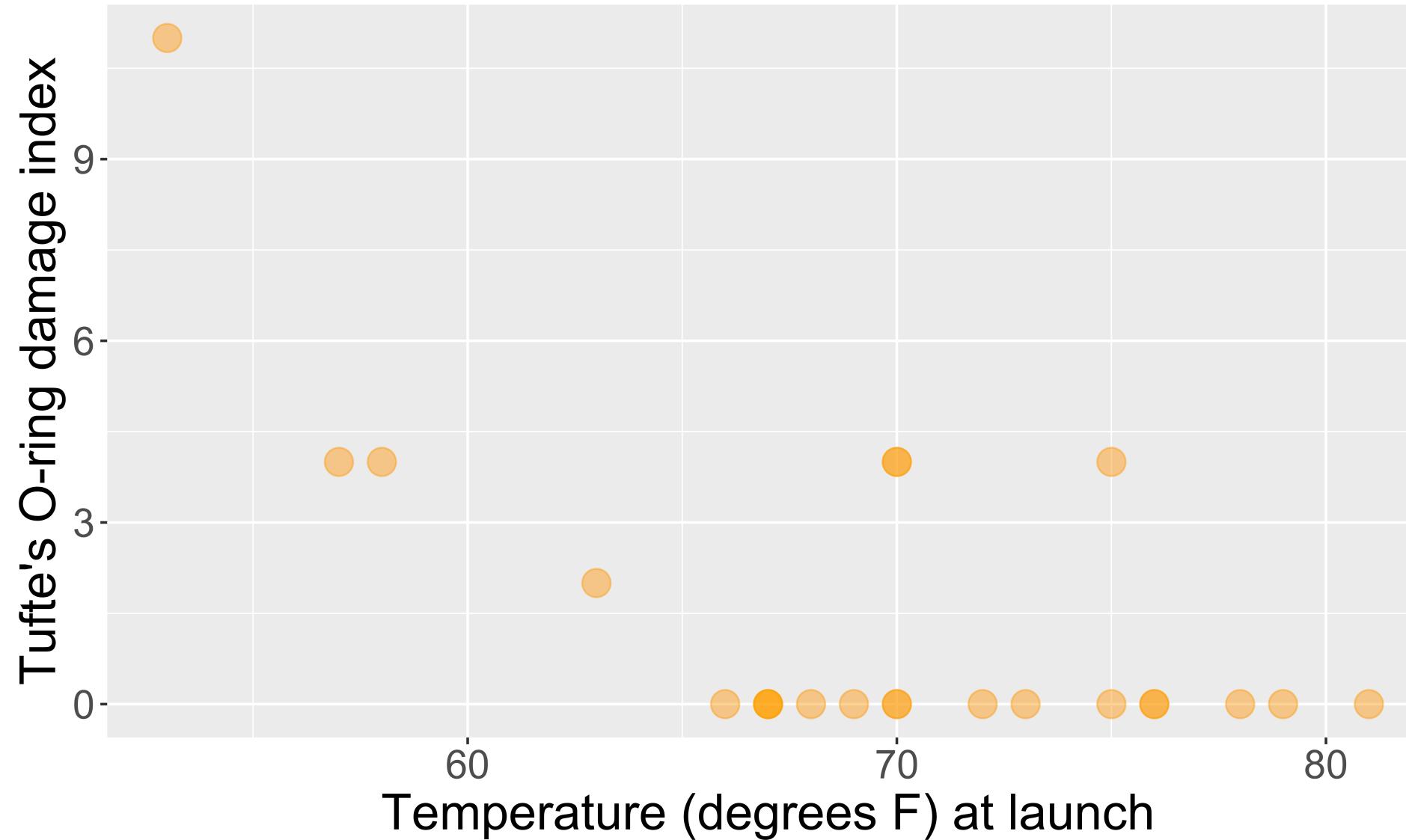
Challenger

Here's another one of those charts.



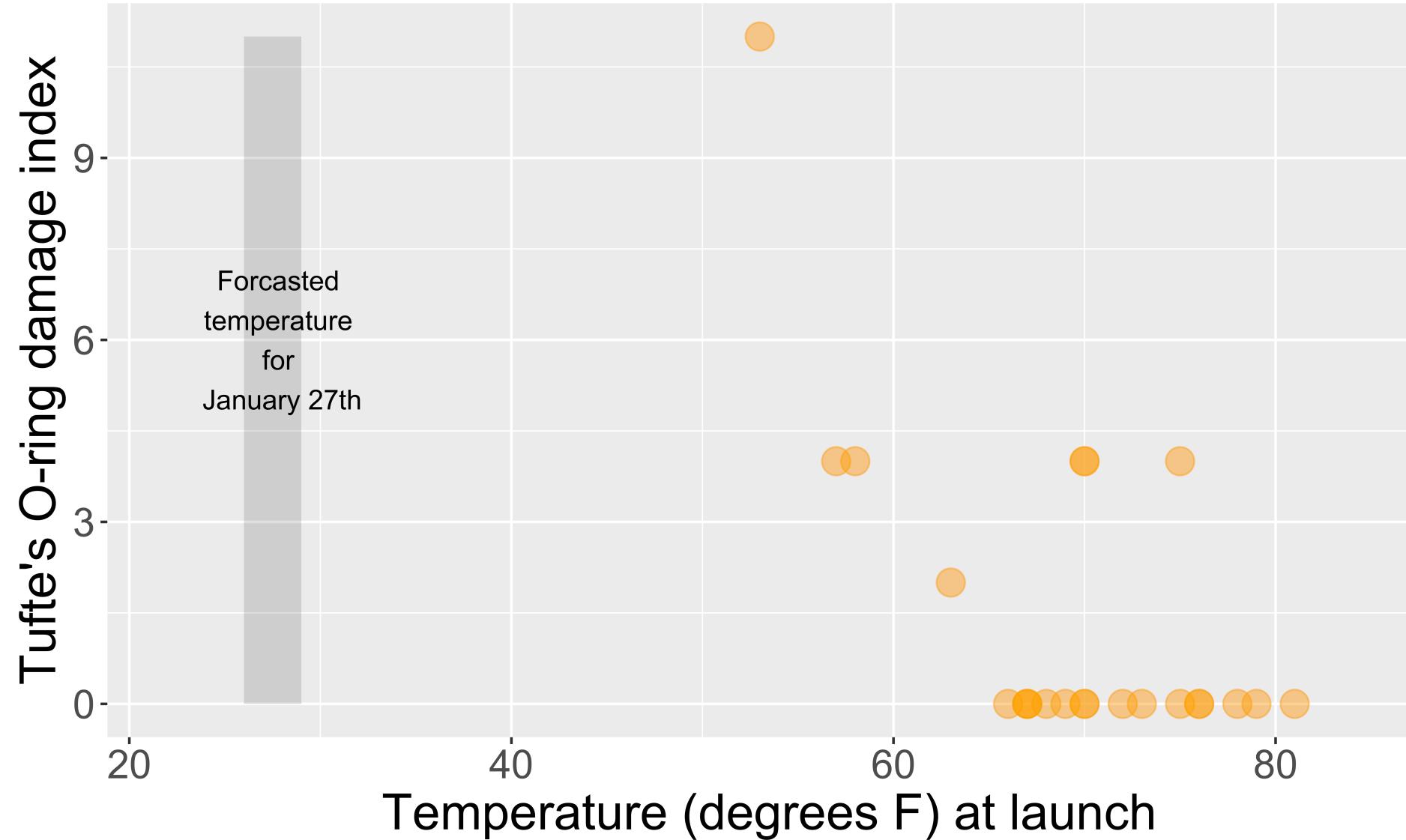
Challenger

Here's a graphic I created from [Edward Tufte's](#) data.



Challenger

This adaptation is a recreation of Edward Tufte's graphic.



Now let's learn the Grammar of Graphics.

We will use this grammar to:

Decompose and understand existing graphs.

Create our own graphs with the R package `ggplot2`.

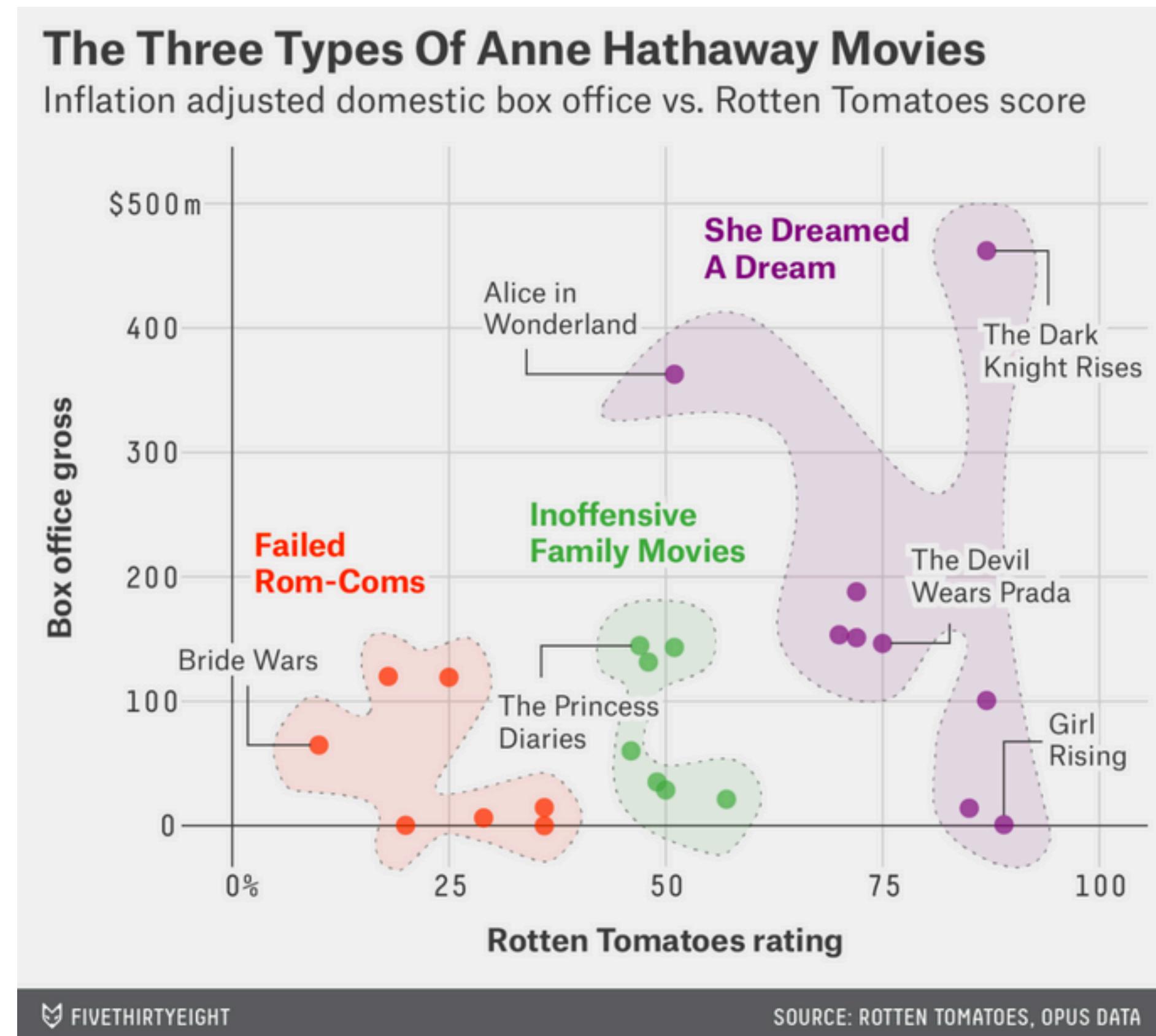
Grammar of Graphics

- **data**: Data frame that contains the raw data
 - Variables used in the graph
- **geom**: Geometric **shape** that the data are mapped to.
 - EX: Point, line, bar, text, ...
- **aesthetic**: Visual properties of the **geom**
 - EX: X (horizontal) position, y (vertical) position, color, fill, shape
- **scale**: Controls how data are mapped to the visual values of the aesthetic.
 - EX: particular colors, log scale
- **guide**: Legend/key to help user convert visual display back to the data

For right now, we won't focus on the **names** of particular types of graphs (e.g., scatterplot) but on the **elements** of graphs.

Example 1

- What are the variables?
- What **geom** are the variables map to?
- What are the **aesthetics** of the **geom**?
- How is each variable mapped to an **aesthetic**?
- What additional context is provided?
- What story is the graph telling?

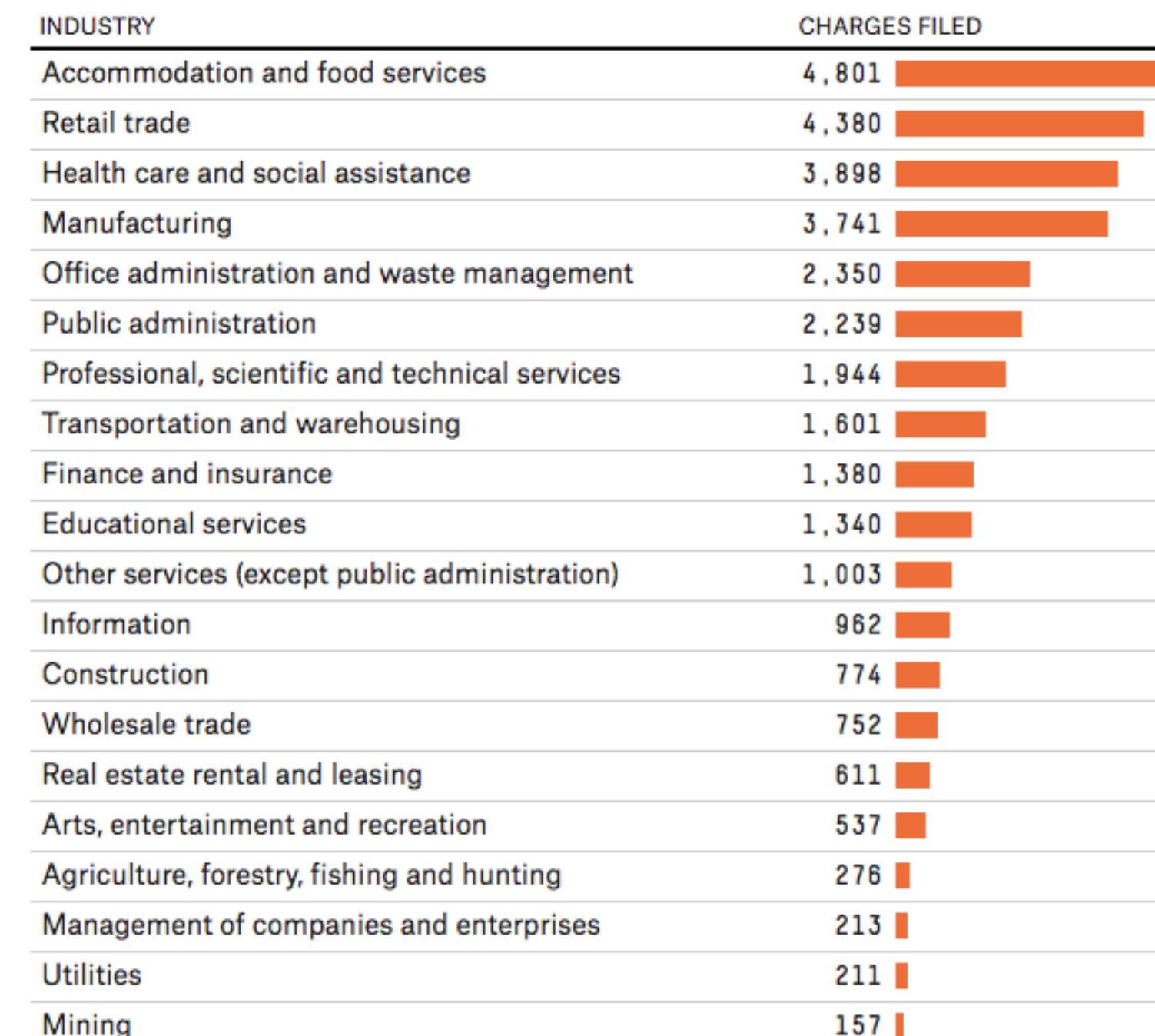


Example 2

- What are the variables?
- What **geom** are the variables map to?
- What are the **aesthetics** of the **geom**?
- How is each variable mapped to an **aesthetic**?
- What additional context is provided?
- What story is the graph telling?

Sexual harassment charges, by industry

Among charges filed by women, fiscal years 2005-2015



Not including 35,304 charges filed without a specified industry

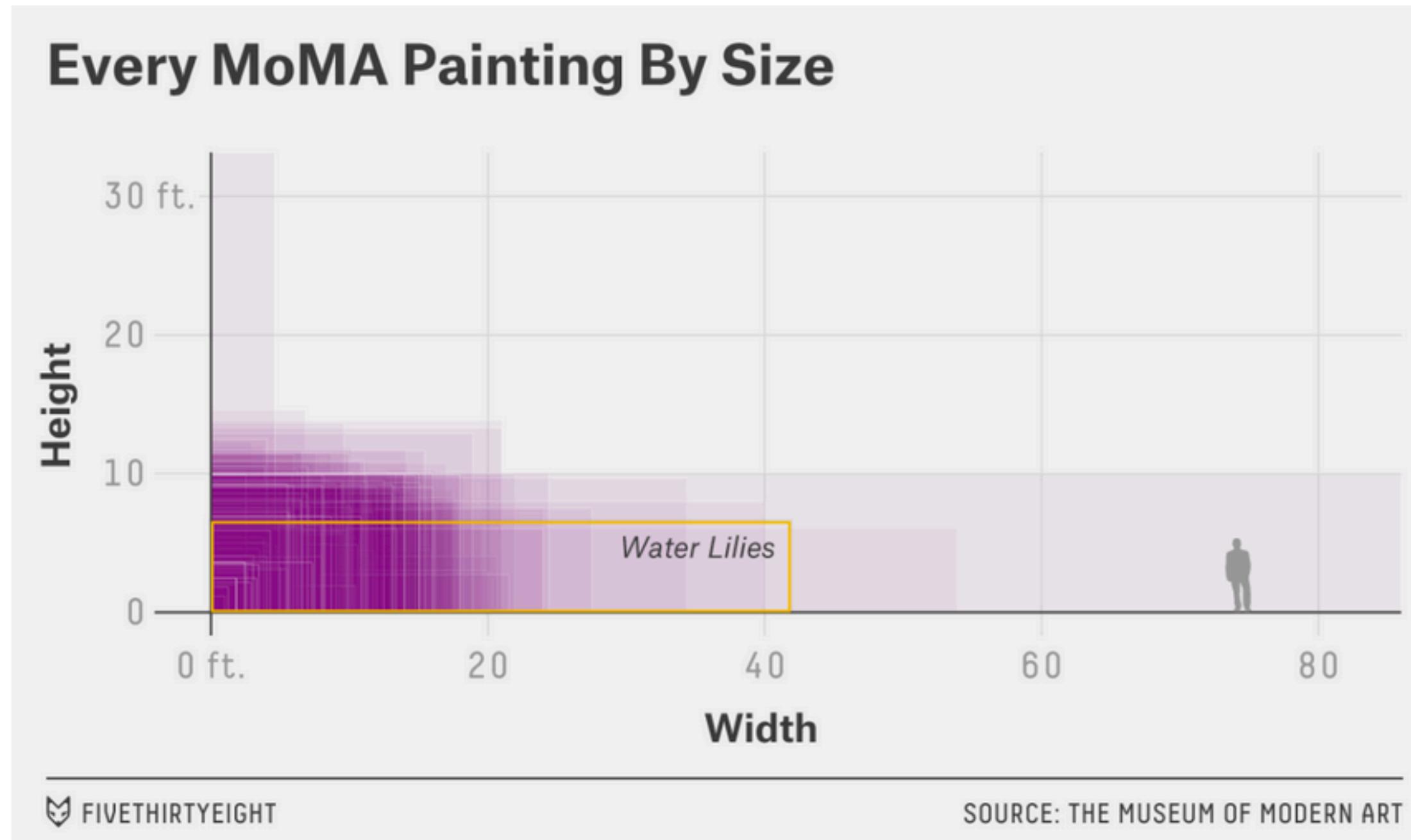
SOURCE: EQUAL EMPLOYMENT OPPORTUNITY COMMISSION

Visualization Considerations

What additional context should my graphs have?

- For context, at a minimum include
 - Axis labels (with units reported).
 - Legends.
 - Data source.
- Think about the **stories/questions** your visualization answers.
- Determine what **context/background information** your viewer needs.
- Visualizing data involves **editorial choices**.
 - What comparisons to make easy to see.
 - What to highlight
 - What references to provide.
 - What scales to use.

Context Example

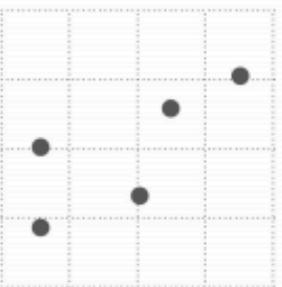


What visual cues are easier to compare?

Visual cues

When you visualize data, you encode values to shapes, sizes, and colors.

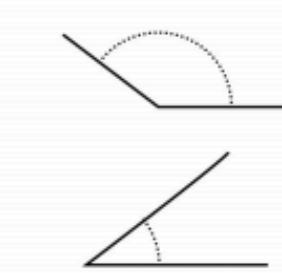
Position
Where in space the data is



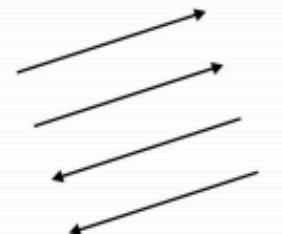
Length
How long the shapes are



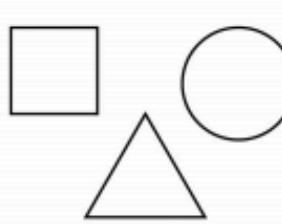
Angle
Rotation between vectors



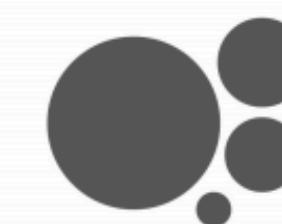
Direction
Slope of a vector in space



Shapes
Symbols as categories



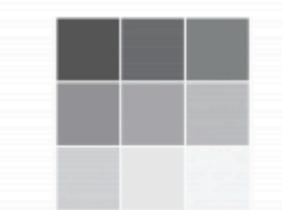
Area
How much 2-D space



Volume
How much 3-D space



Color saturation
Intensity of a color hue



Color hue
Usually referred to as color

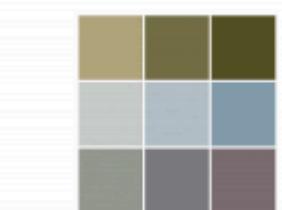
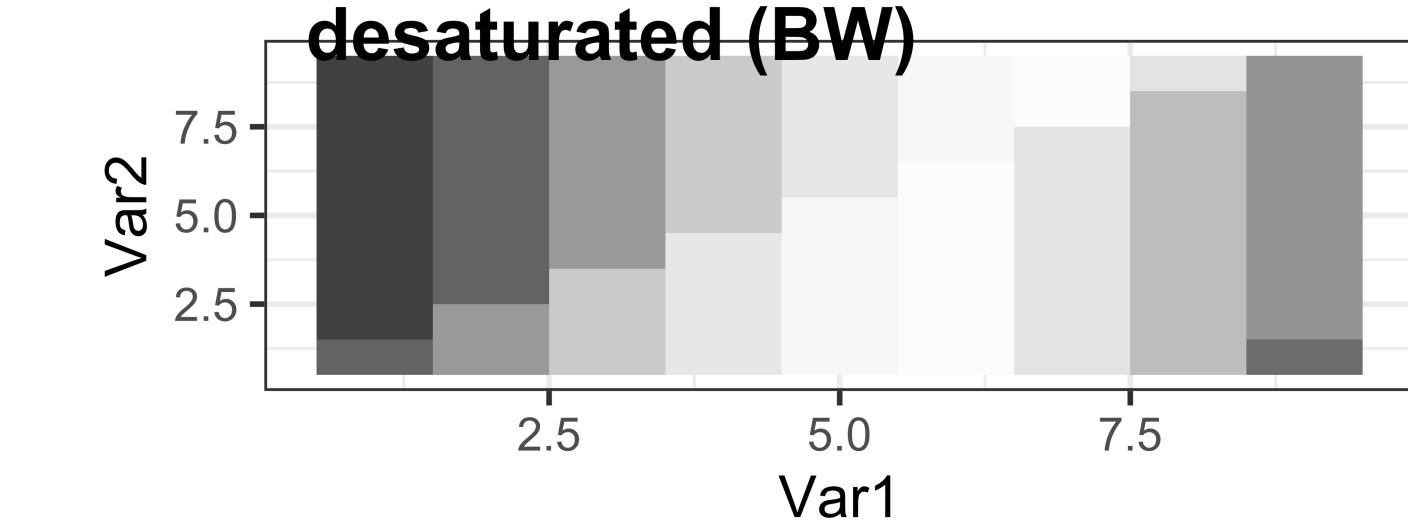
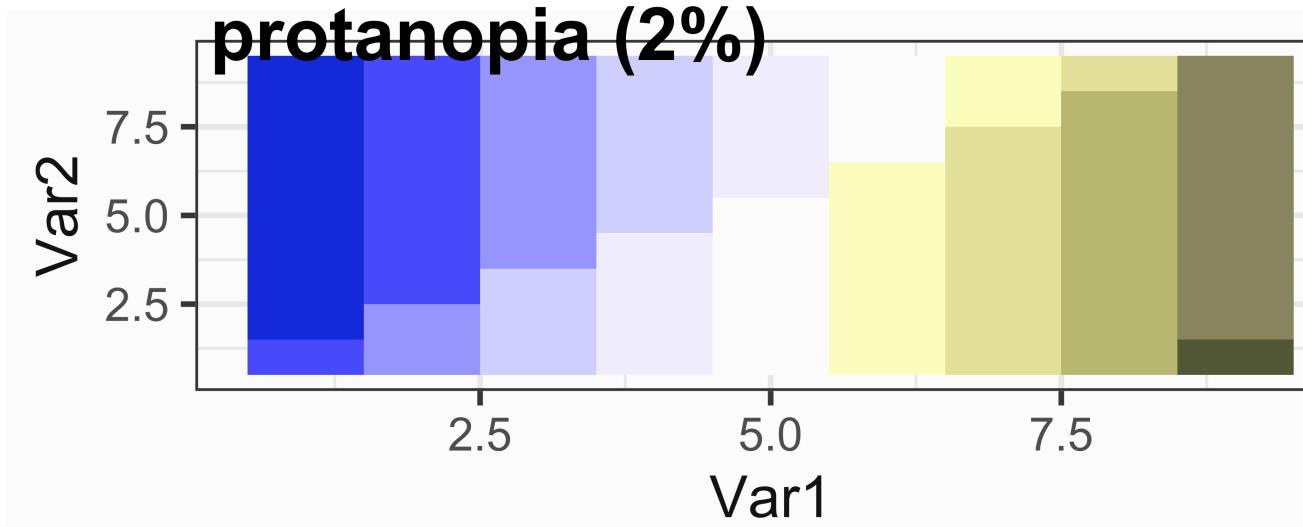
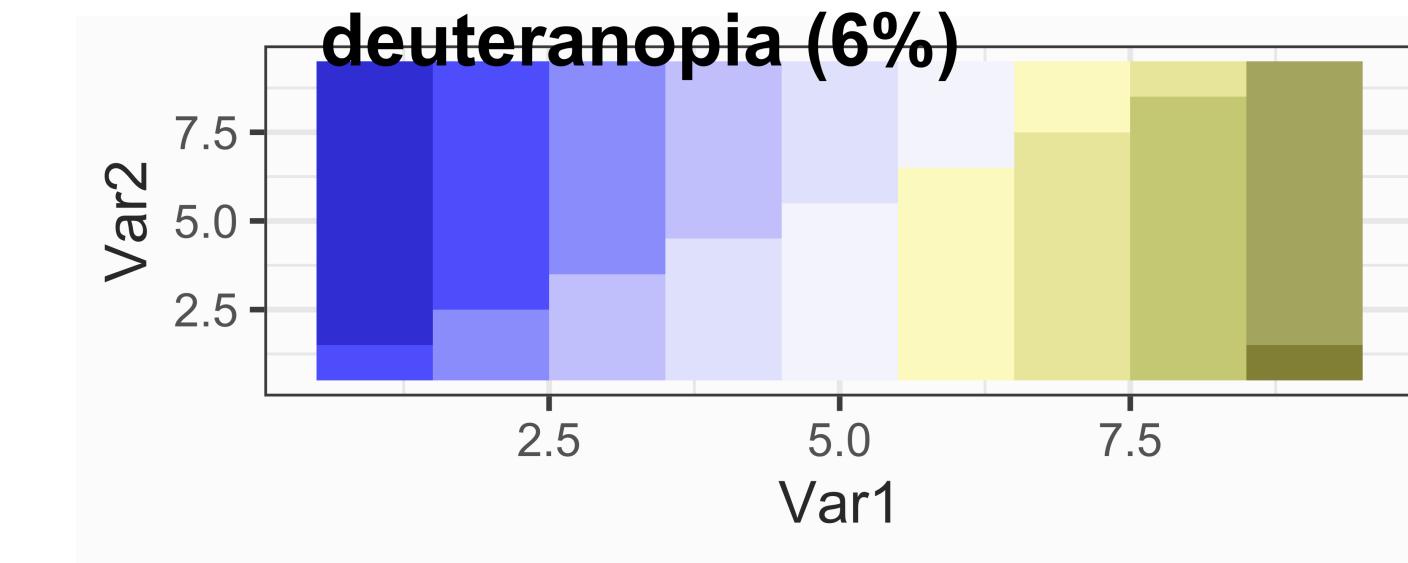
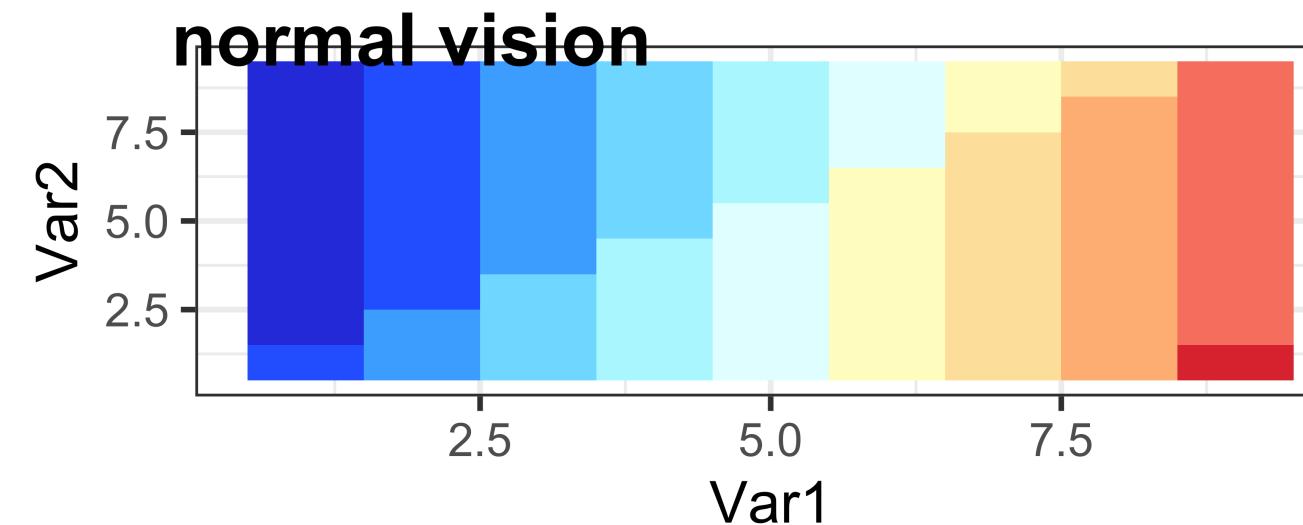


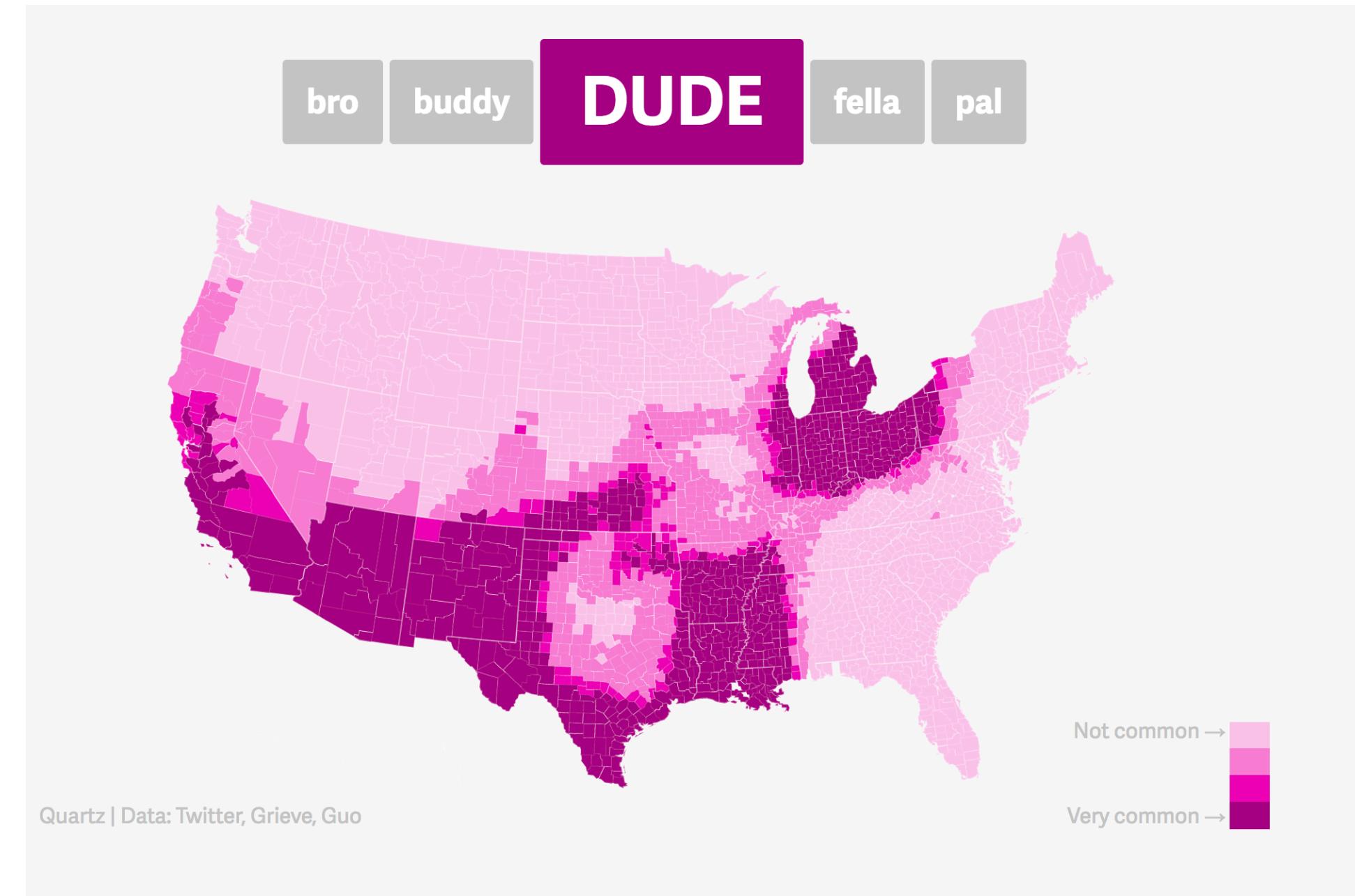
FIGURE 3-3 Visual cues

What to consider with color?

Consider color blindness.

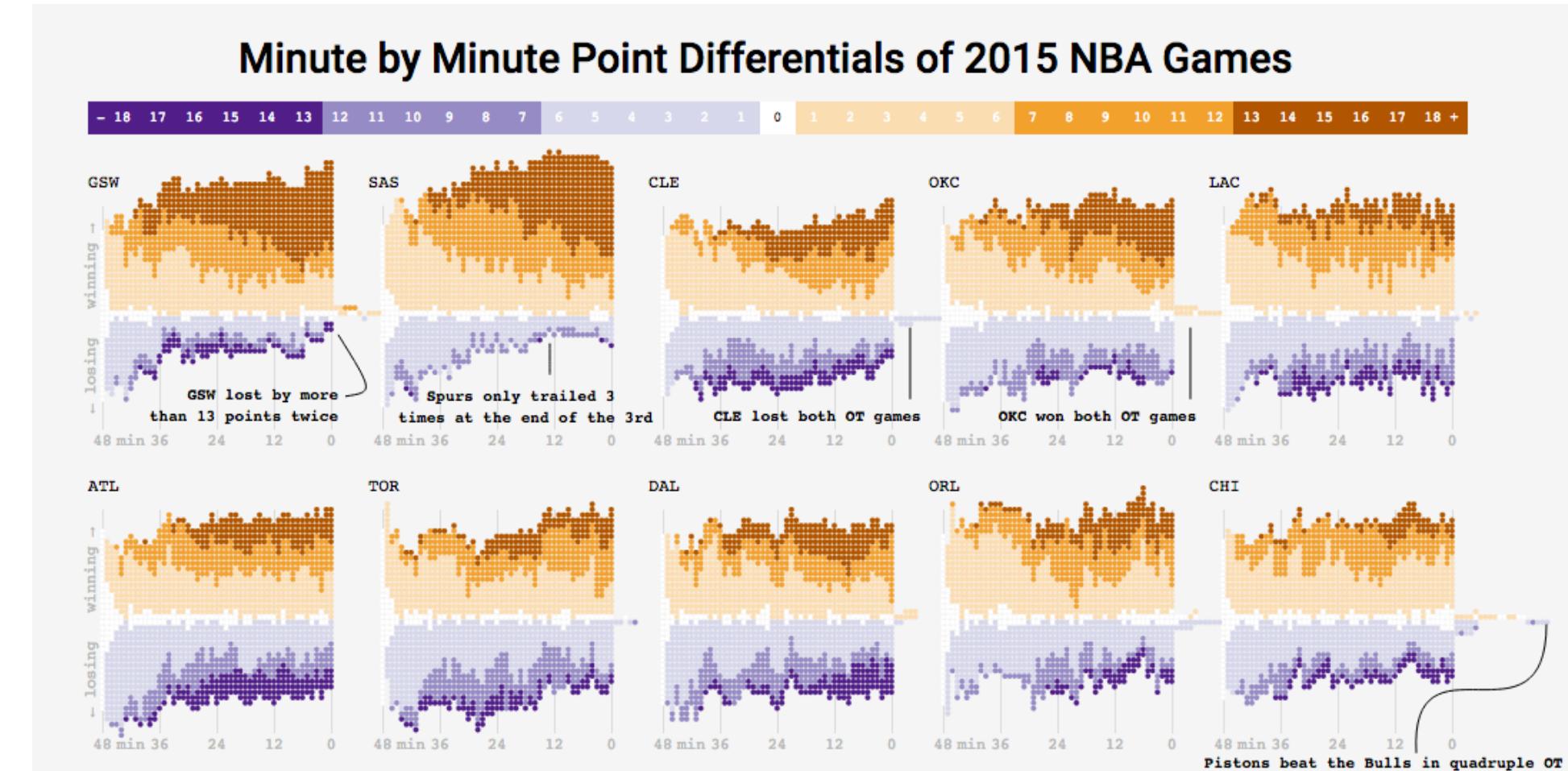


Color Palettes – Sequential



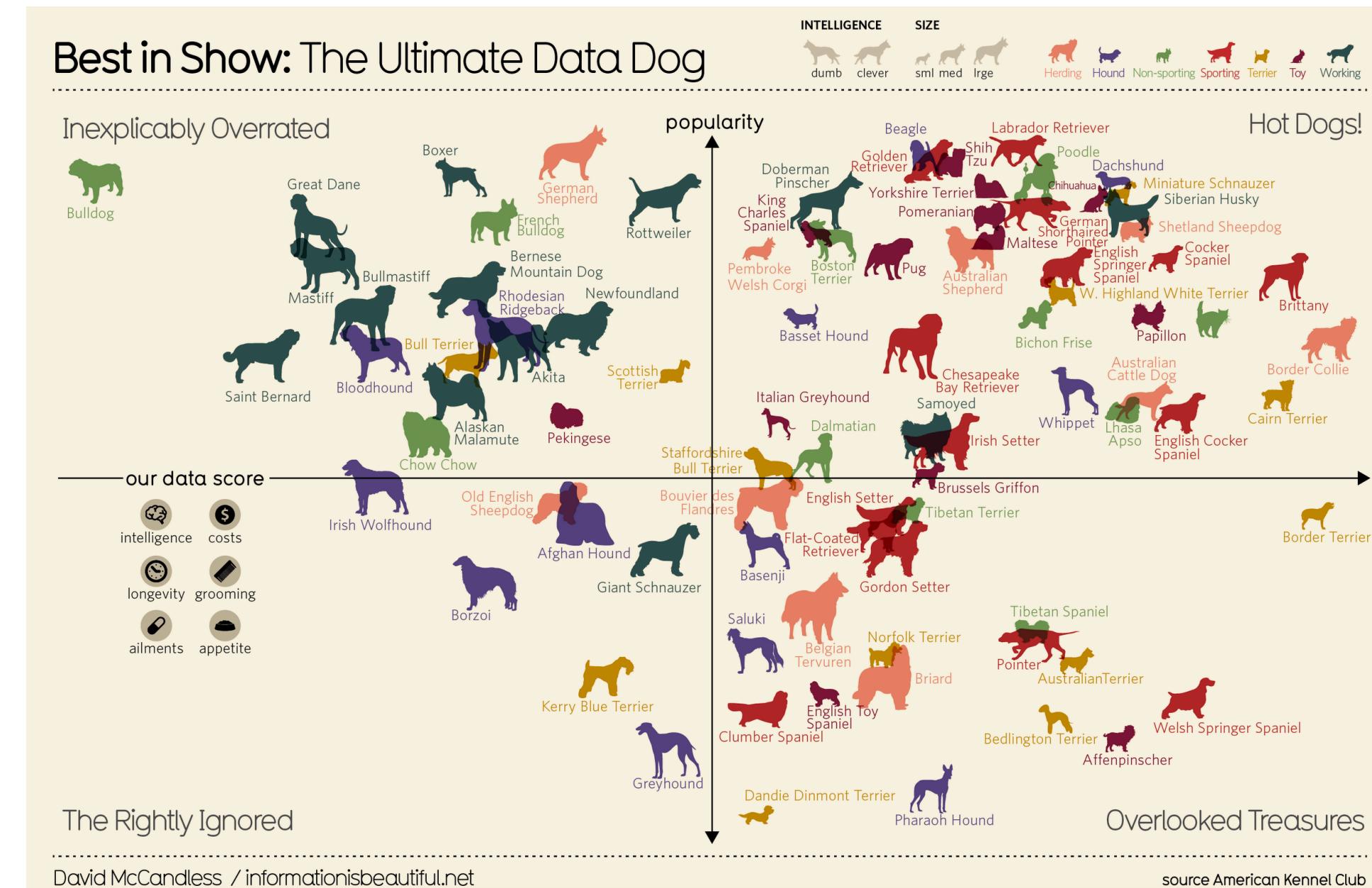
Maps, like the [Dude map](#) are also a great way to provide context!

Color Palettes – Diverging



Adam Pearce's 2015 NBA Games

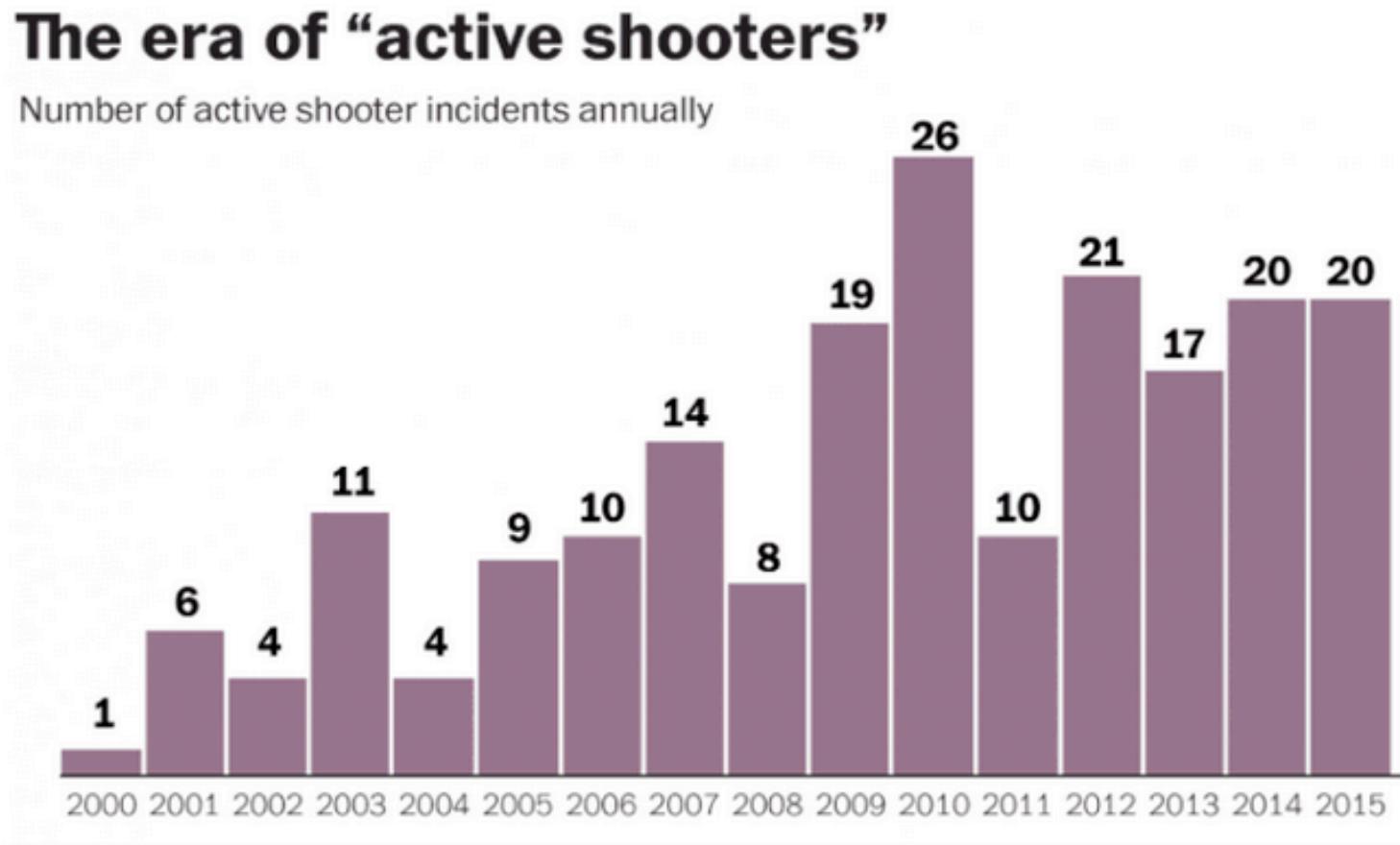
Color Palettes – Qualitative



information is beautiful's Best in Show

Many Ways To Visually Tell A Story

Washington Post's Approach:

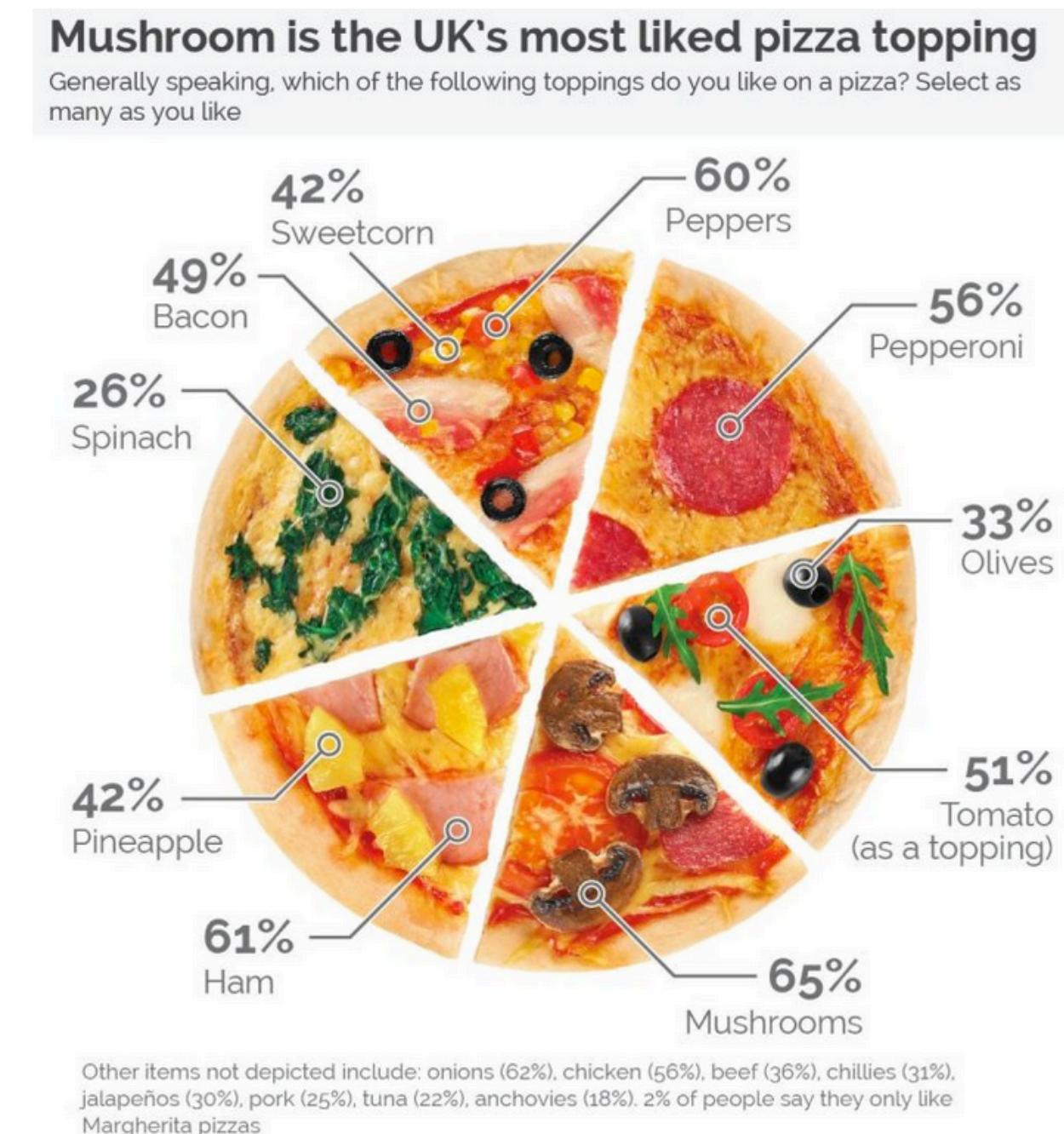


A bar chart of the number of “active shooter” incidents in the United States between 2000 and 2015.
Credit: *The Washington Post* WonkBlog

Periscopic's Approach

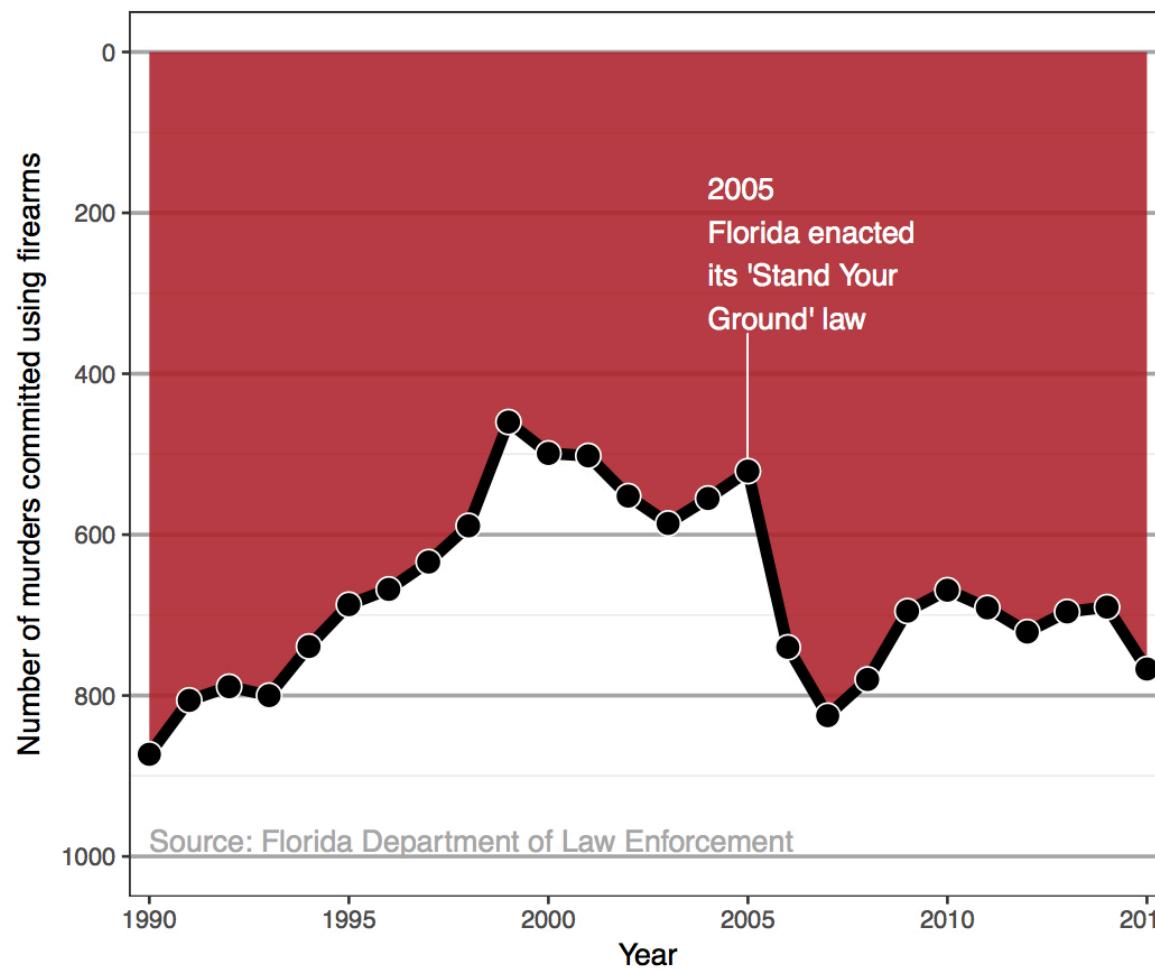
Bad Graphics

Because of all the design choices, it is much easier to make a bad graph than a good graph.



Misleading Graphics

Be careful that your design choices don't cause your viewer to draw incorrect conclusions about the data:

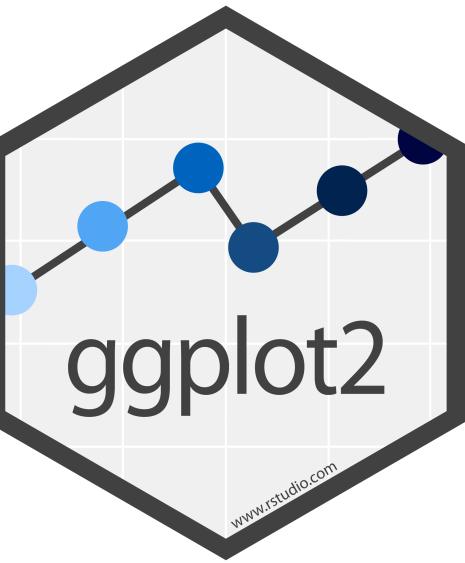


- Just letting the software make all the design choices can still lead to misleading graphs (recall the Georgia COVID graph).

Summary Thoughts on Graphical Considerations

- Good graphics are ones where the findings and insights are **obvious** to the viewer.
 - Add information and key **context**.
- Facilitate the **comparisons** that correspond to the research question.
 - Recall the three Georgia COVID counts graphs from Day 1!
- Data visualizations are **not neutral**.
- It is easier to see the differences and similarities between different types of graphics if we learn the **grammar of graphics**.
- Practicing **decomposing** graphics should make it easier for us to **compose** our own graphics.

Load Necessary Packages



ggplot2 is part of this collection of data science packages.

```
1 # Load necessary packages  
2 library(tidyverse)
```

Data Setting: Eco-Totem Broadway Bicycle Count



Import the Data

```
1 july_2019 <- read_csv("data/july_2019.csv")
2
3 # Inspect the data
4 glimpse(july_2019)
```

```
Rows: 192
Columns: 8
$ DateTime <chr> "07/04/2019 12:00:00 AM", "07/04/2019 12:15:00 AM", "07/04/2...
$ Day      <chr> "Thursday", "Thursday", "Thursday", "Thursday", ...
$ Date     <date> 2019-07-04, 2019-07-04, 2019-07-04, 2019-07-04, ...
$ Time     <time> 00:00:00, 00:15:00, 00:30:00, 00:45:00, 01:00:00, 01:15:00, ...
$ Total    <dbl> 2, 3, 2, 0, 3, 2, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, ...
$ Westbound <dbl> 2, 3, 1, 0, 2, 2, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, ...
$ Eastbound <dbl> 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, ...
$ Occasion <chr> "Fourth of July", "Fourth of July", "Fourth of July", "Four...
```

Inspect the Data

```
1 # Look at first few rows  
2 head(july_2019)
```

# A tibble: 6 × 8								
	DateTime	Day	Date	Time	Total	Westbound	Eastbound	Occasion
	<chr>	<chr>	<date>	<tim>	<dbl>	<dbl>	<dbl>	<chr>
1	07/04/2019 12:00:00...	Thur...	2019-07-04	00:00	2	2	0	Fourth ...
2	07/04/2019 12:15:00...	Thur...	2019-07-04	00:15	3	3	0	Fourth ...
3	07/04/2019 12:30:00...	Thur...	2019-07-04	00:30	2	1	1	Fourth ...
4	07/04/2019 12:45:00...	Thur...	2019-07-04	00:45	0	0	0	Fourth ...
5	07/04/2019 01:00:00...	Thur...	2019-07-04	01:00	3	2	1	Fourth ...
6	07/04/2019 01:15:00...	Thur...	2019-07-04	01:15	2	2	0	Fourth ...

What does a row represent here?

Inspect the Data

```
1 # Determine type  
2 # To access one variable: dataset$variable  
3 class(july_2019$Day)  
  
[1] "character"  
  
1 class(july_2019$Total)  
  
[1] "numeric"  
  
1 class(july_2019)  
  
[1] "spec_tbl_df" "tbl_df"        "tbl"          "data.frame"
```

Grammar of Graphics

- **data**: Data frame that contains the raw data
 - Variables used in the graph
- **geom**: Geometric **shape** that the data are mapped to.
 - EX: Point, line, bar, text, ...
- **aesthetic**: Visual properties of the **geom**
 - EX: X (horizontal) position, y (vertical) position, color, fill, shape
- **scale**: Controls how data are mapped to the visual values of the aesthetic.
 - EX: particular colors, log scale
- **guide**: Legend/key to help user convert visual display back to the data

ggplot2 example code

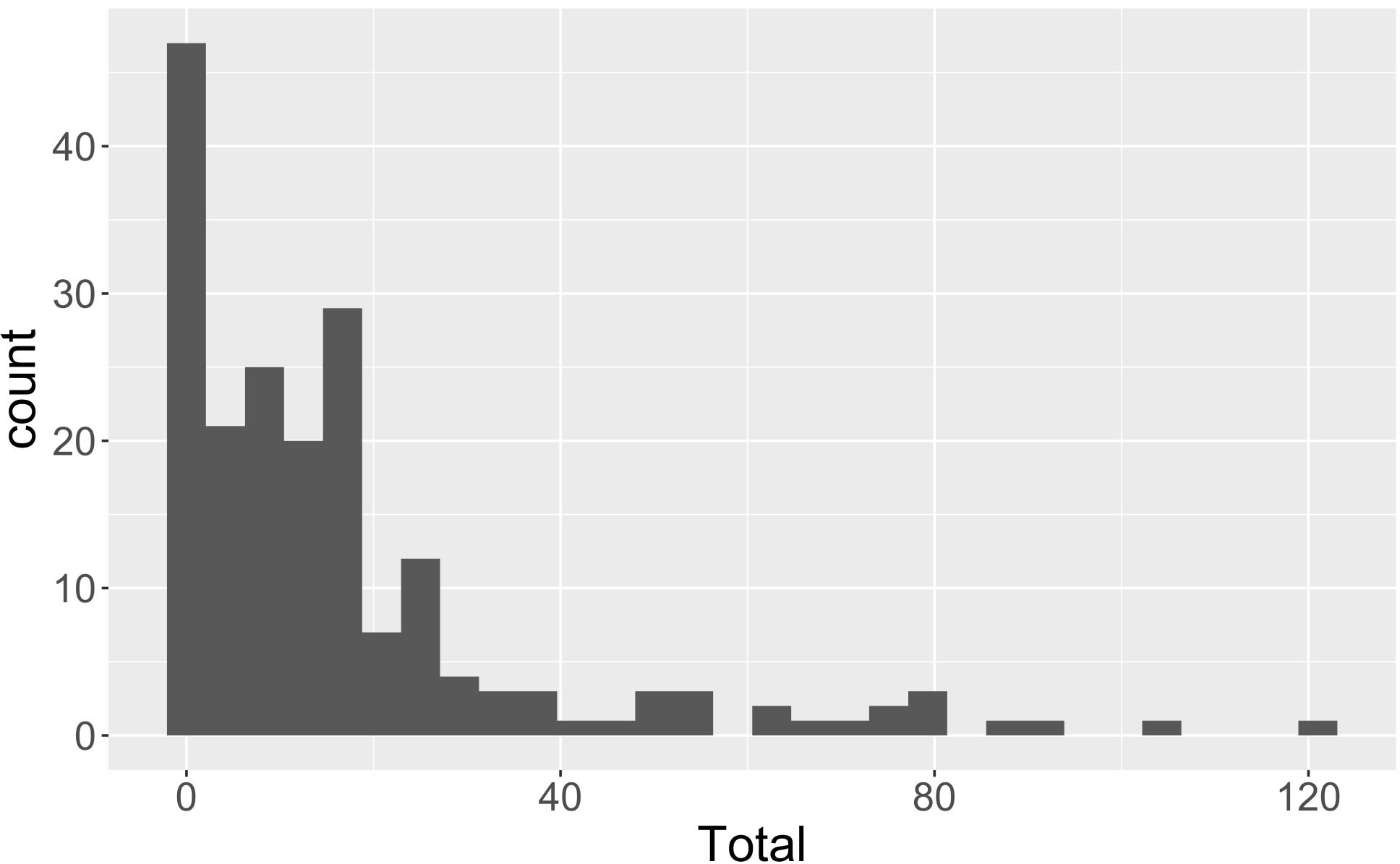
Guiding Principle: We will map variables from the **data** to the **aesthetic** attributes (e.g. location, size, shape, color) of **geometric** objects (e.g. points, lines, bars).

```
1 ggplot(data = ---, mapping = aes(---)) +  
2   geom_---(---)
```

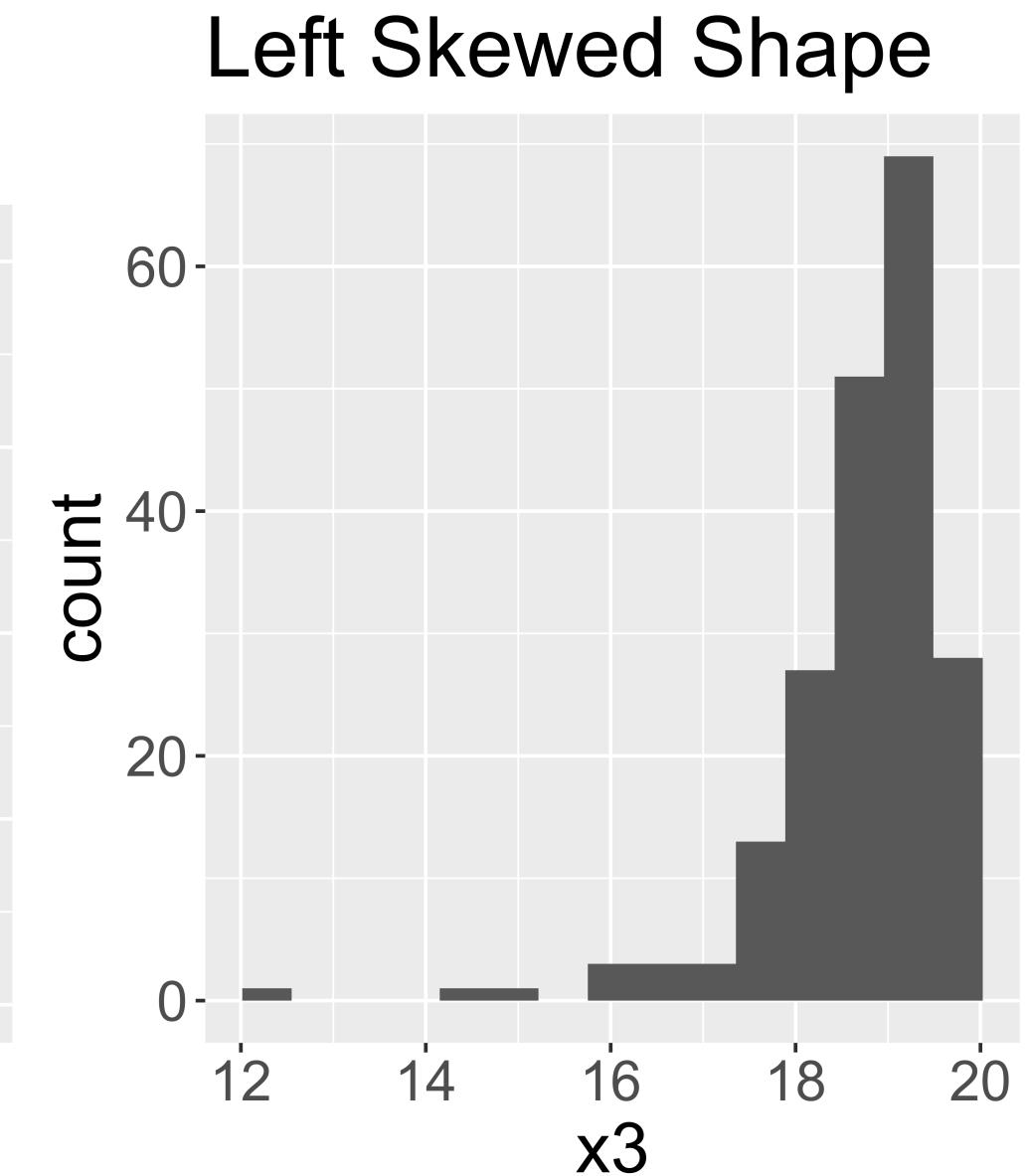
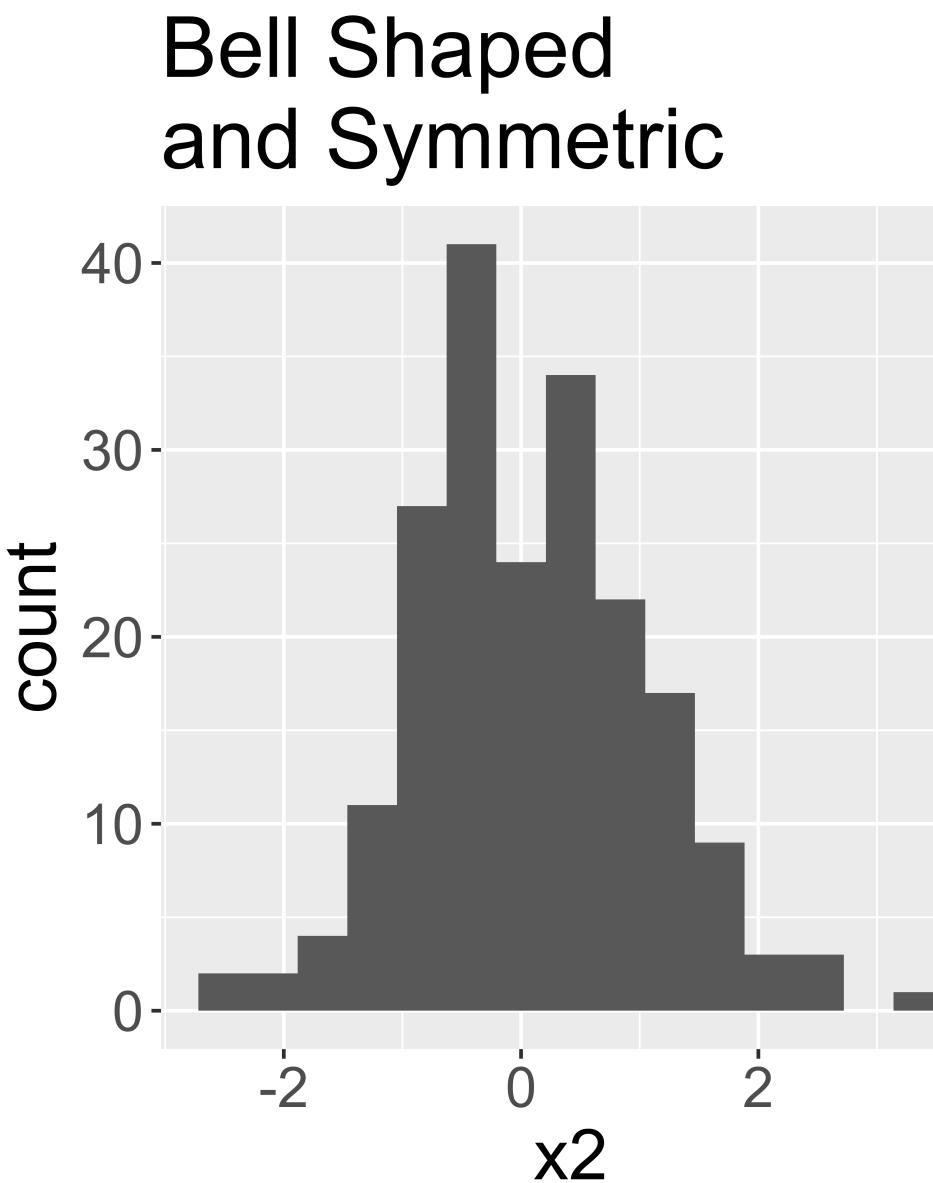
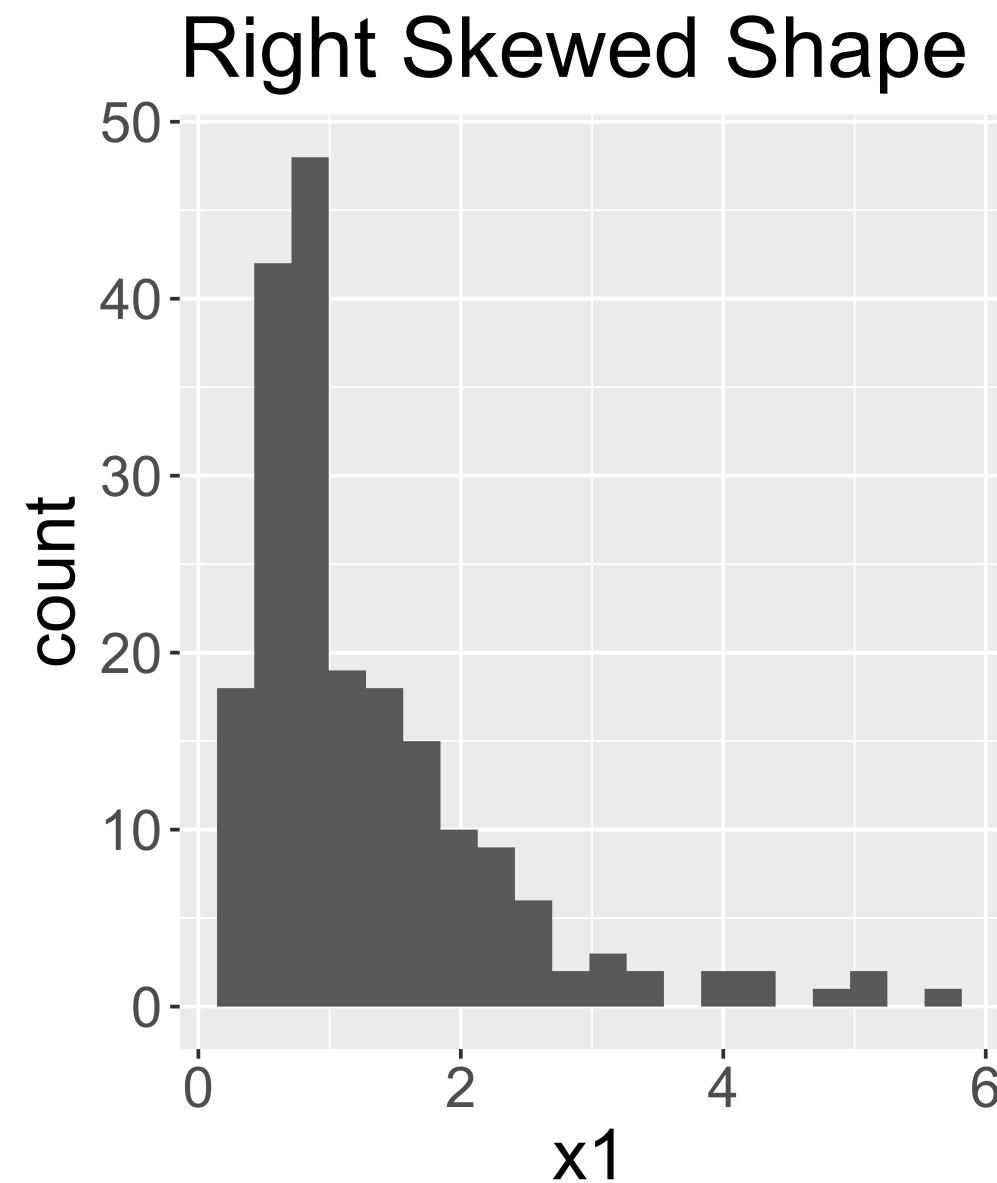
- There are other layers, such as **scales_----_()** and **labs()**, but we will wait on those.

Histograms

- Binned counts of data.
- Great for assessing shape.

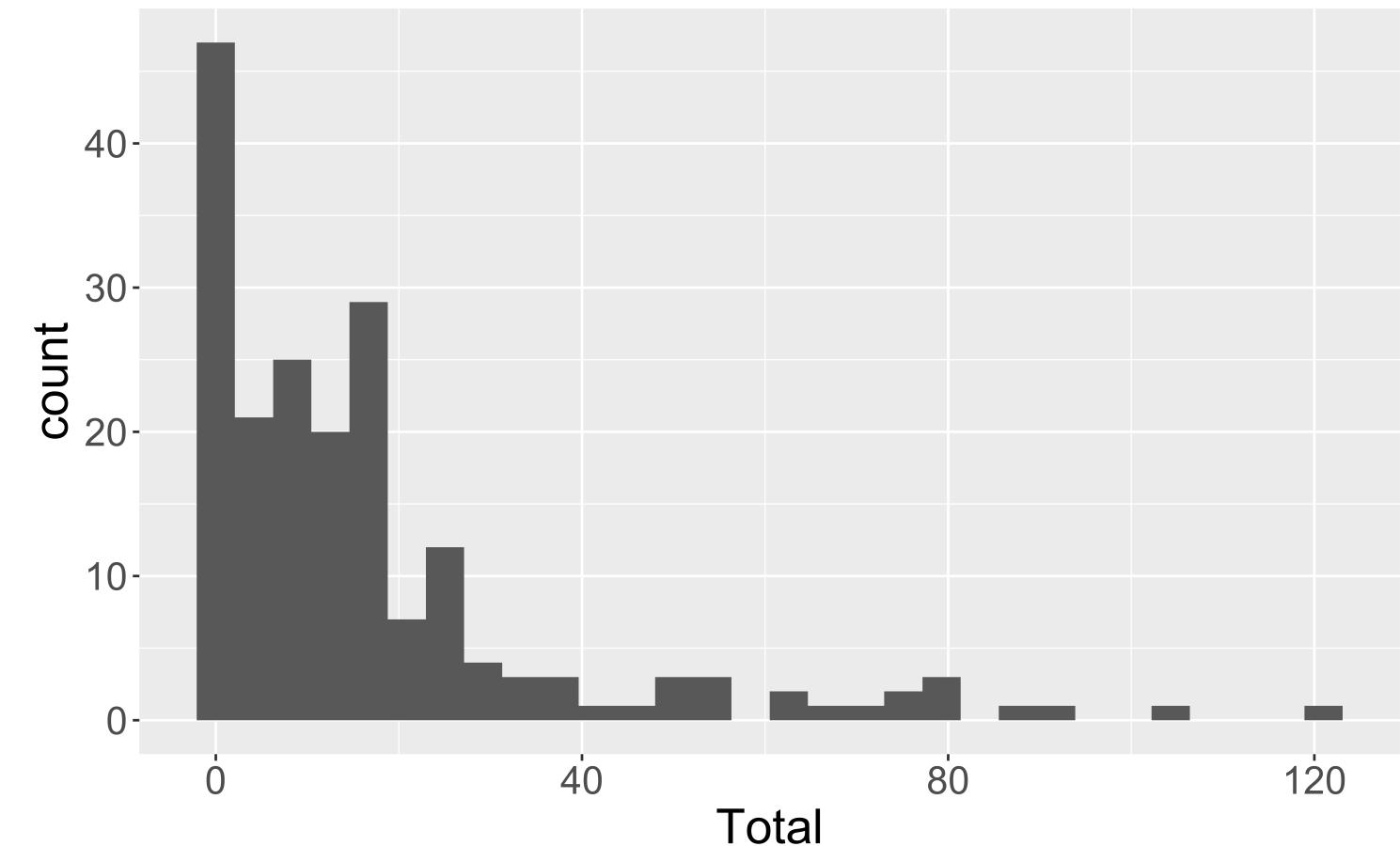


Data Shapes



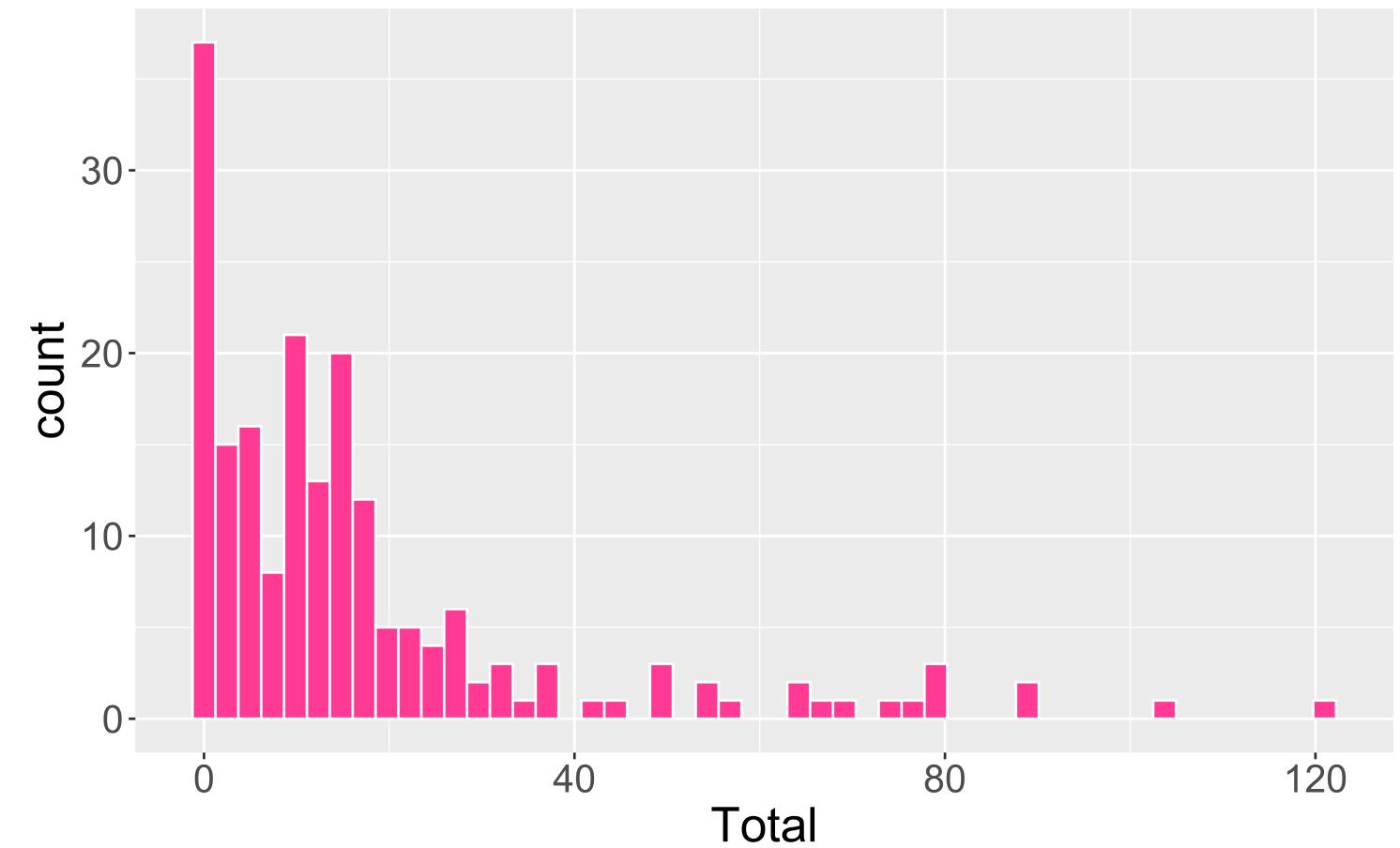
Histograms

```
1 # Create histogram
2 ggplot(data = july_2019,
3         mapping = aes(x = Total)) +
4   geom_histogram()
```



Histograms

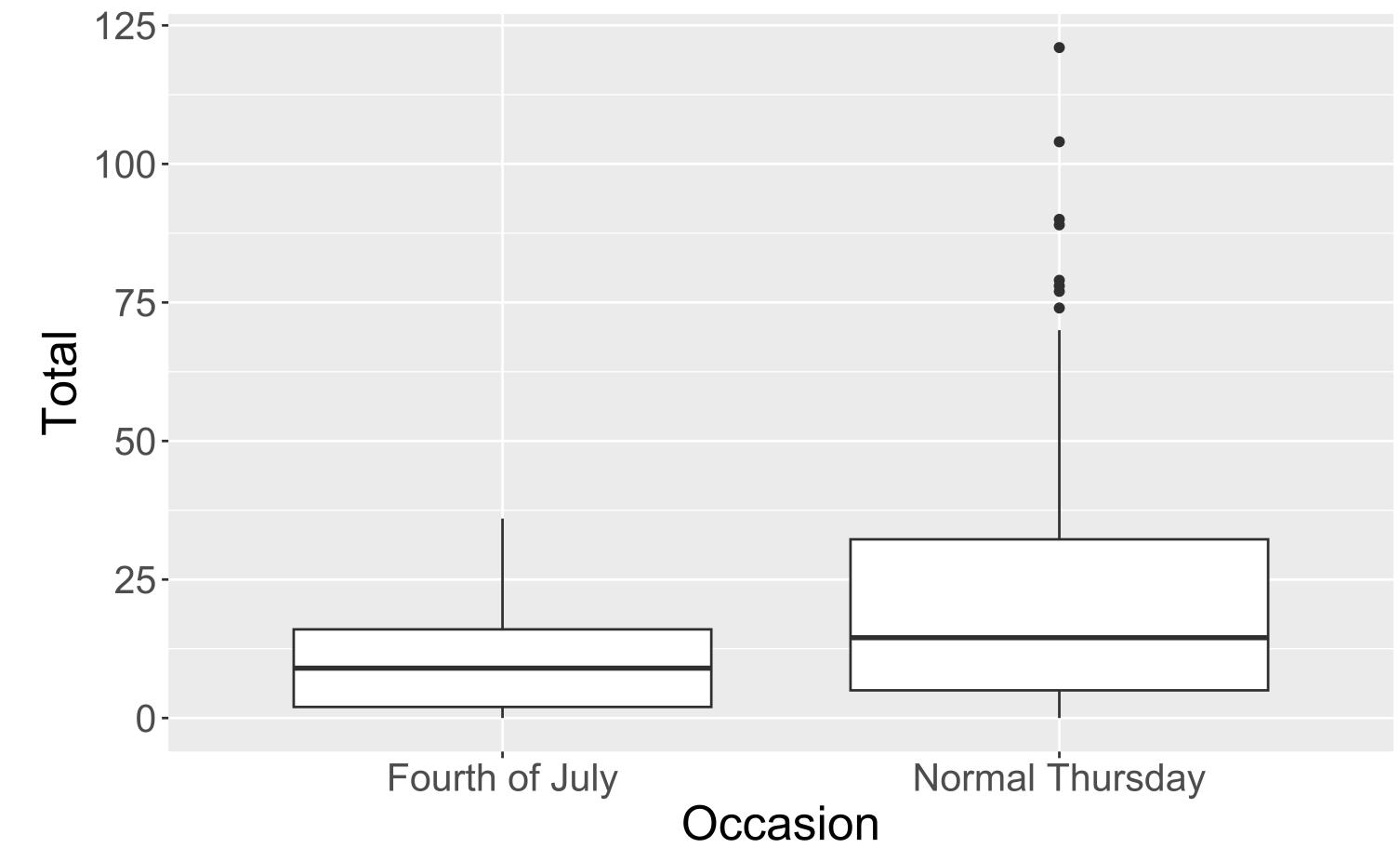
```
1 # Create histogram
2 ggplot(data = july_2019,
3         mapping = aes(x = Total)) +
4         geom_histogram(color = "white",
5                         fill = "violetred1",
6                         bins = 50)
```



- **mapping** to a variable goes in **aes()**
- **setting** to a specific value goes in the **geom_---**()

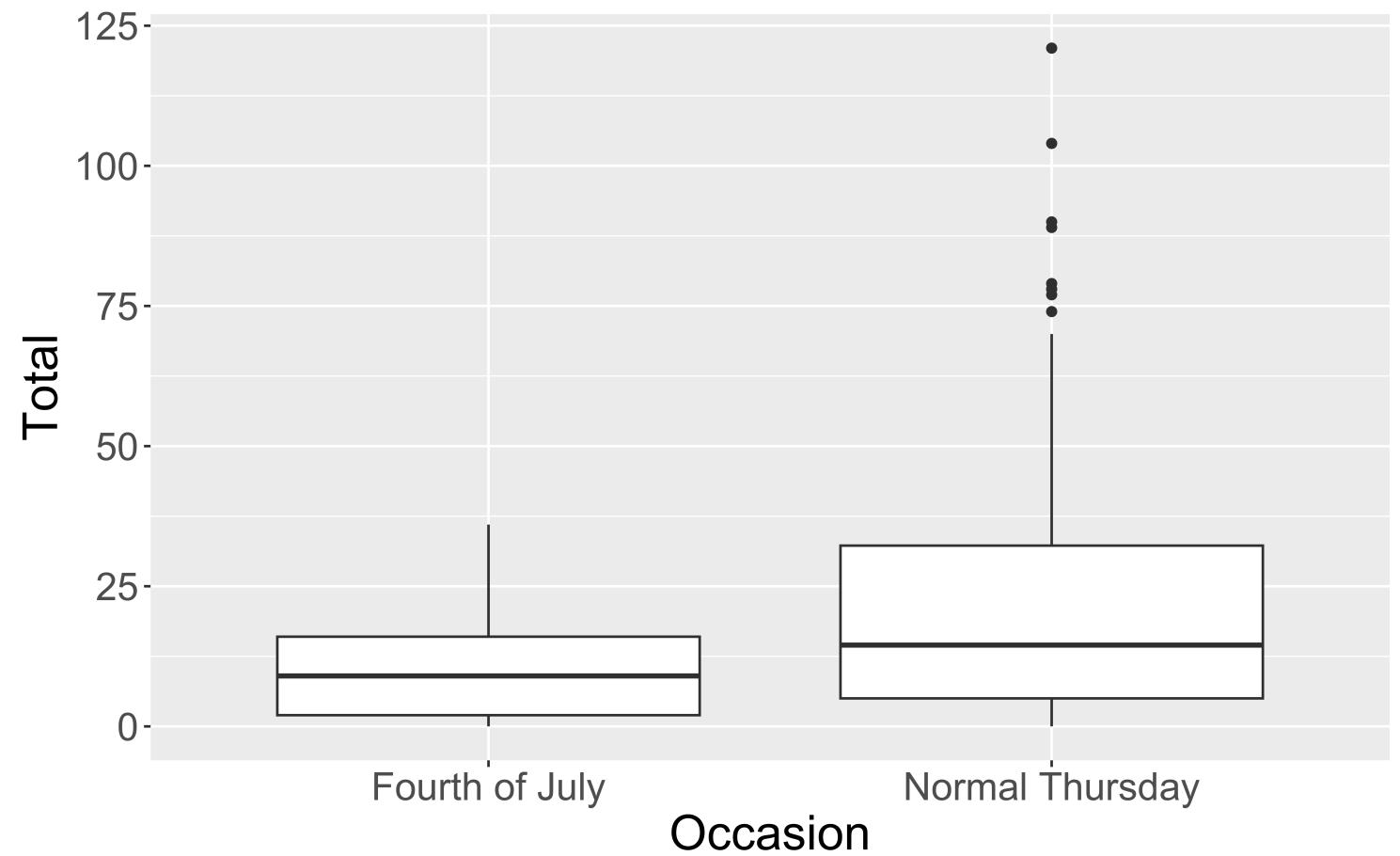
Boxplots

- **Five number summary:**
 - Minimum
 - First quartile (Q1)
 - Median
 - Third quartile (Q3)
 - Maximum
- Interquartile range (IQR) = $Q3 - Q1$
- Outliers: **unusual** points
 - Boxplot defines unusual as being beyond $1.5 * IQR$ from $Q1$ or $Q3$.
- Whiskers: reach out to the furthest point that is NOT an outlier



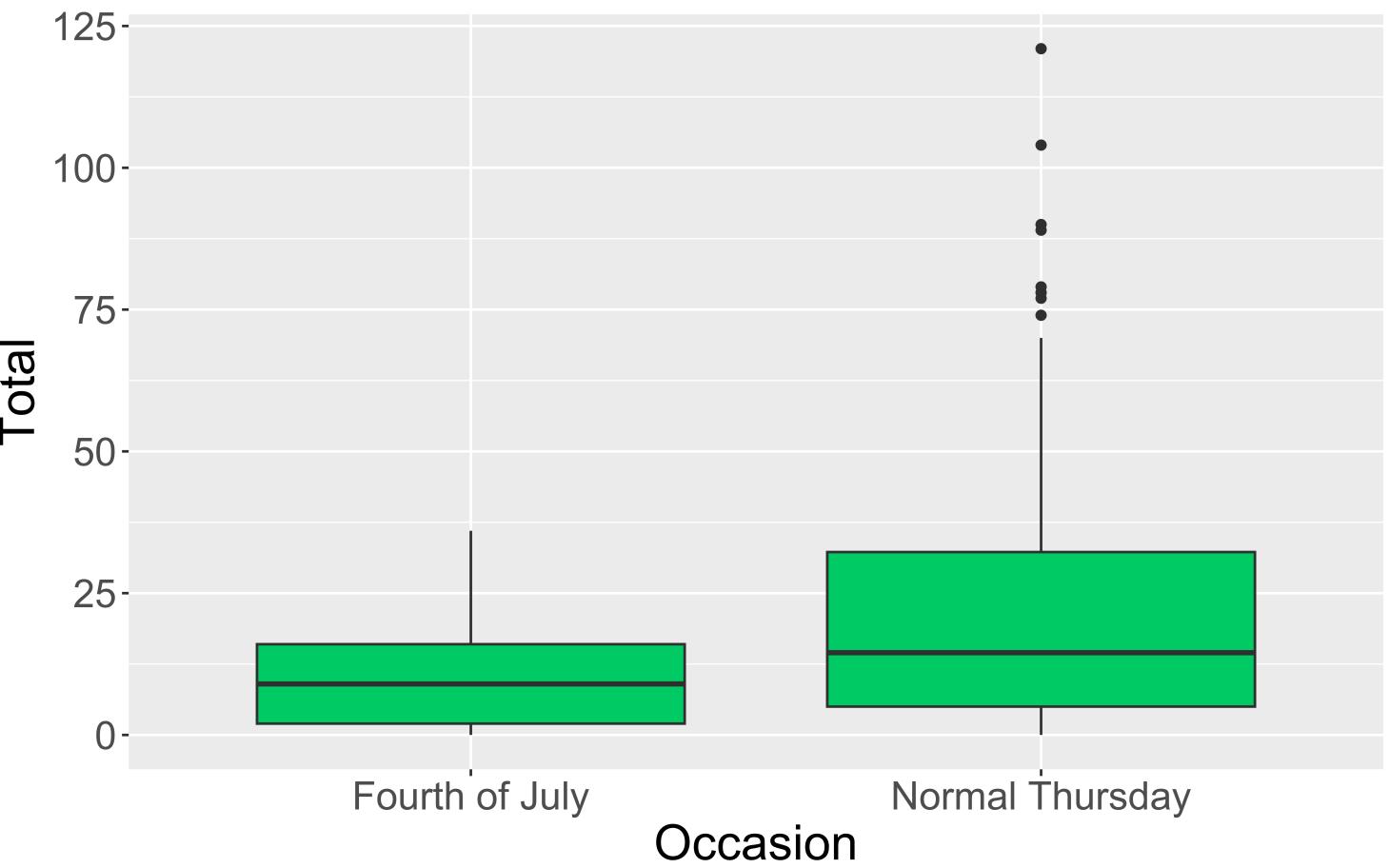
Boxplots

```
1 # Create boxplot
2 ggplot(data = july_2019,
3         mapping = aes(x = Occasion,
4                           y = Total)) +
5   geom_boxplot()
```



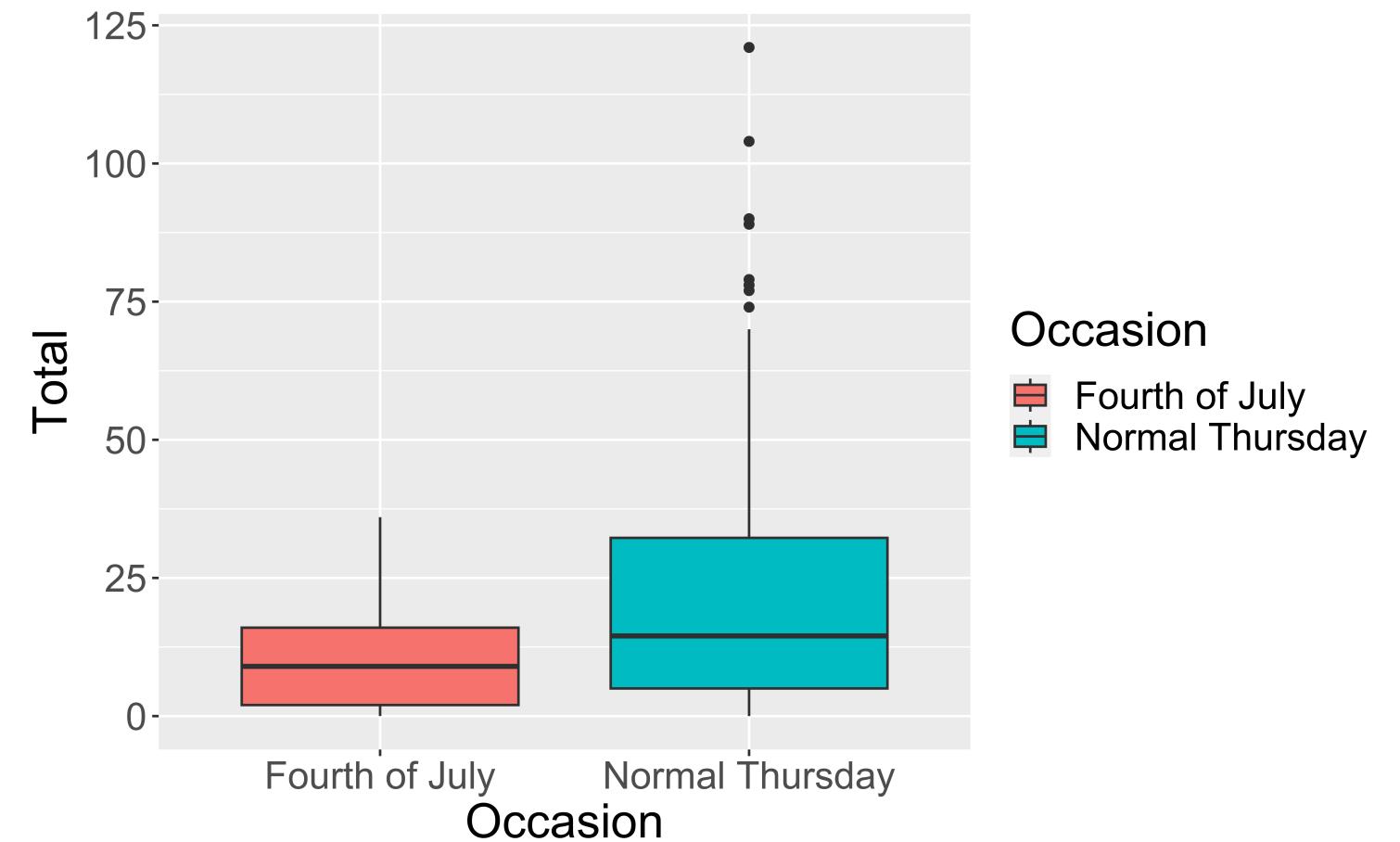
Boxplots

```
1 ggplot(data = july_2019,  
2         mapping = aes(x = Occasion,  
3                           y = Total)) +  
4         geom_boxplot(fill = "springgreen3")
```



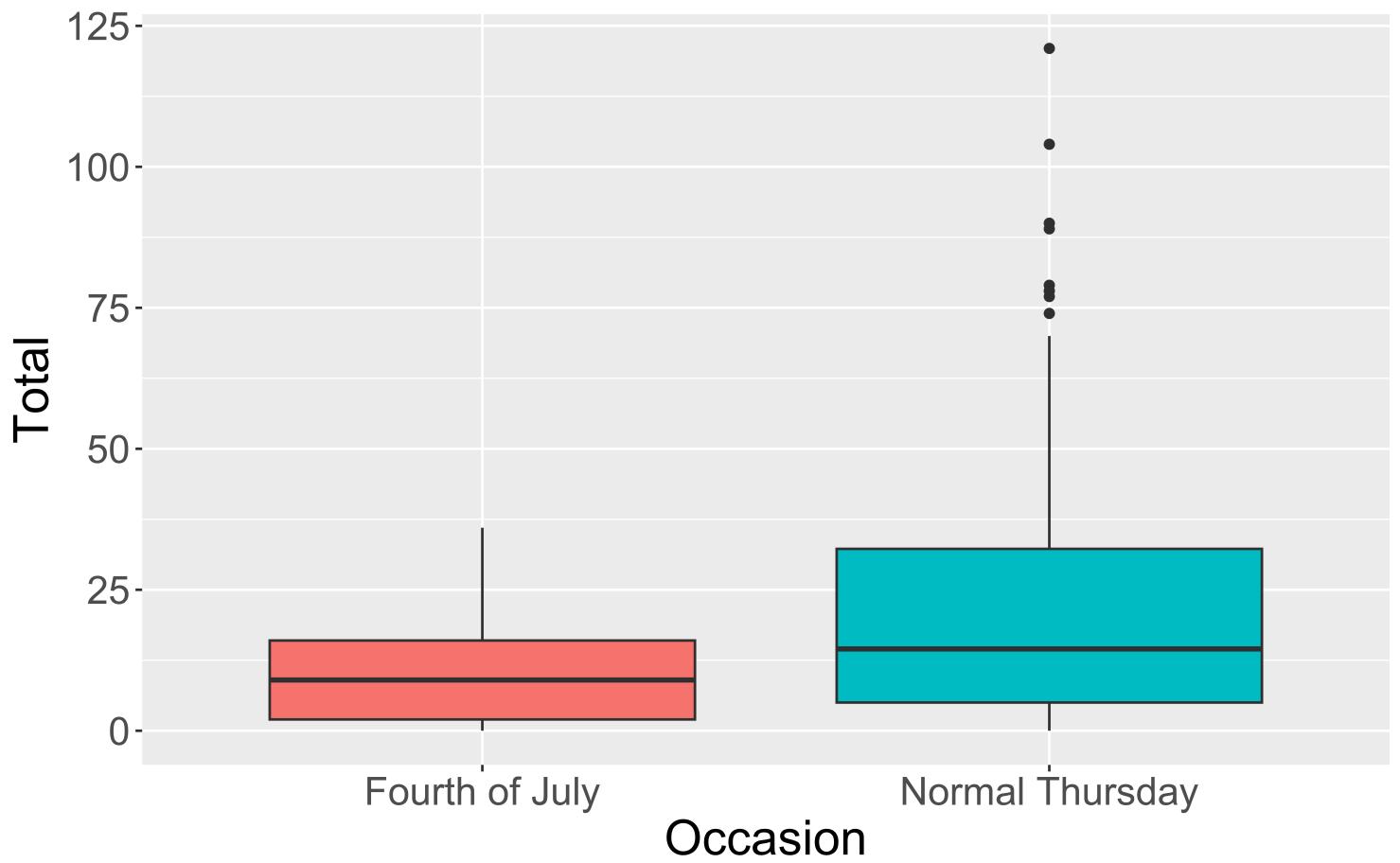
Boxplots

```
1 ggplot(data = july_2019,  
2         mapping = aes(x = Occasion,  
3                           y = Total,  
4                           fill = Occasion)) +  
5   geom_boxplot()
```



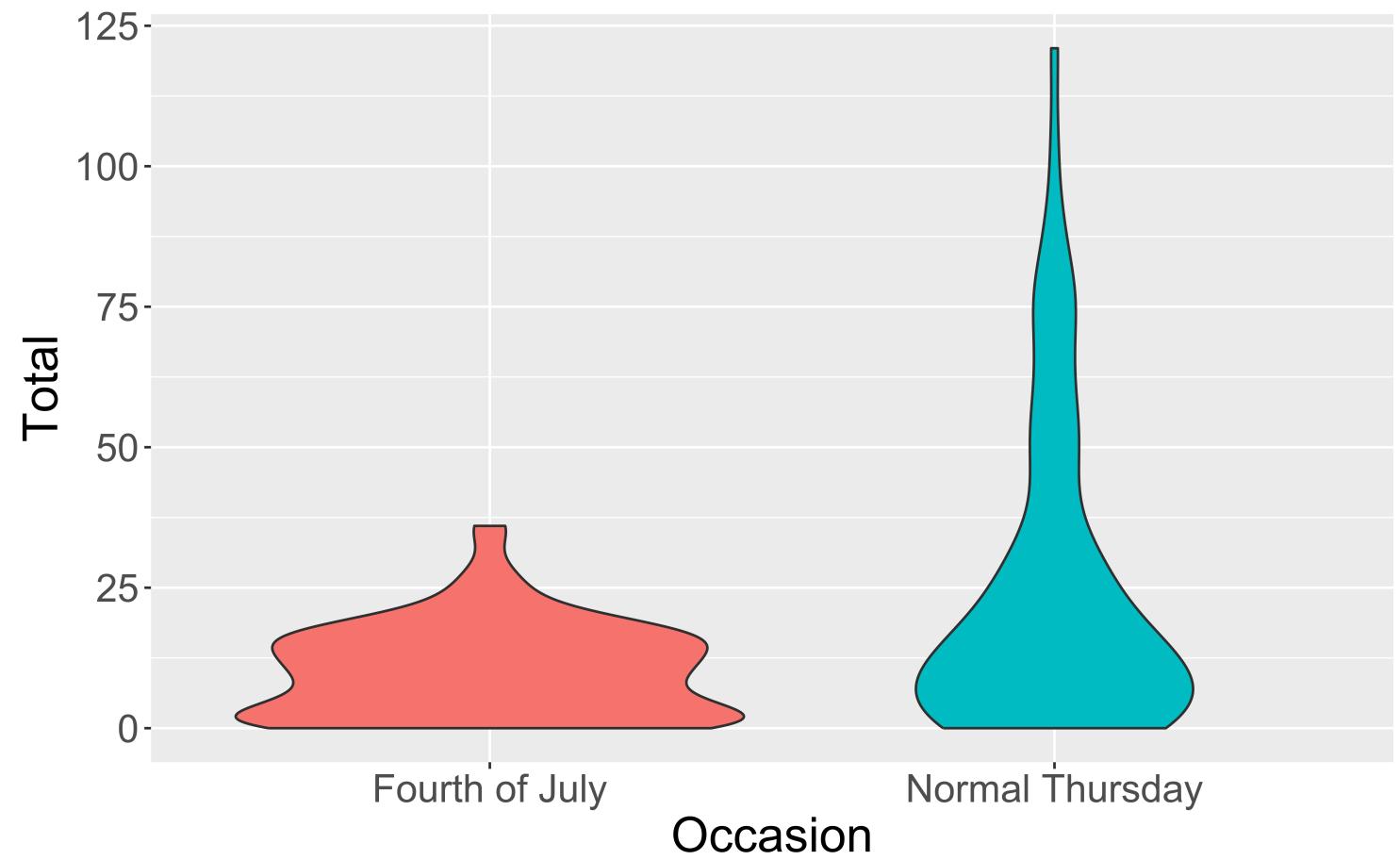
Boxplots

```
1 ggplot(data = july_2019,  
2         mapping = aes(x = Occasion,  
3                           y = Total,  
4                           fill = Occasion)) +  
5   geom_boxplot() +  
6   guides(fill = "none")
```

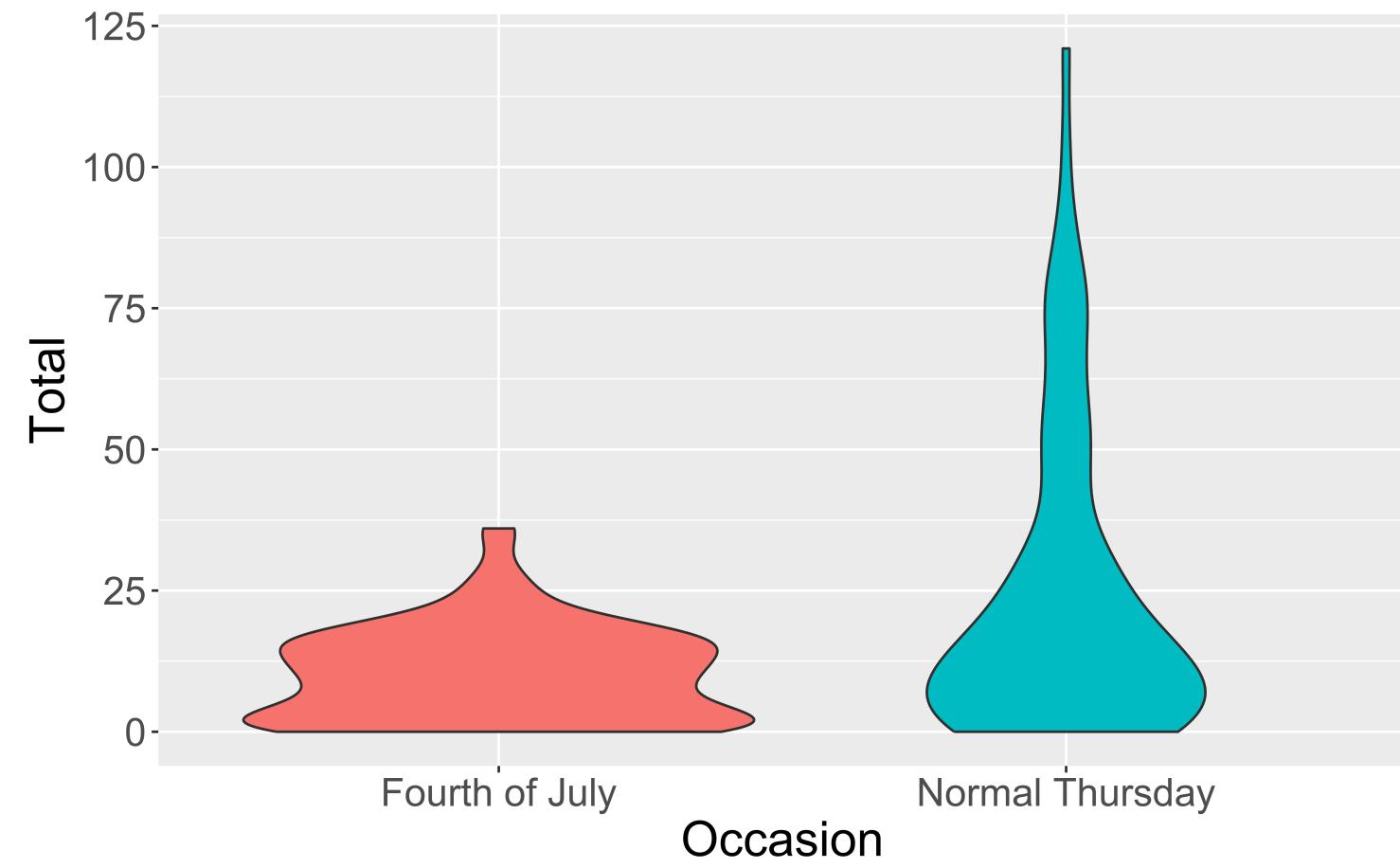
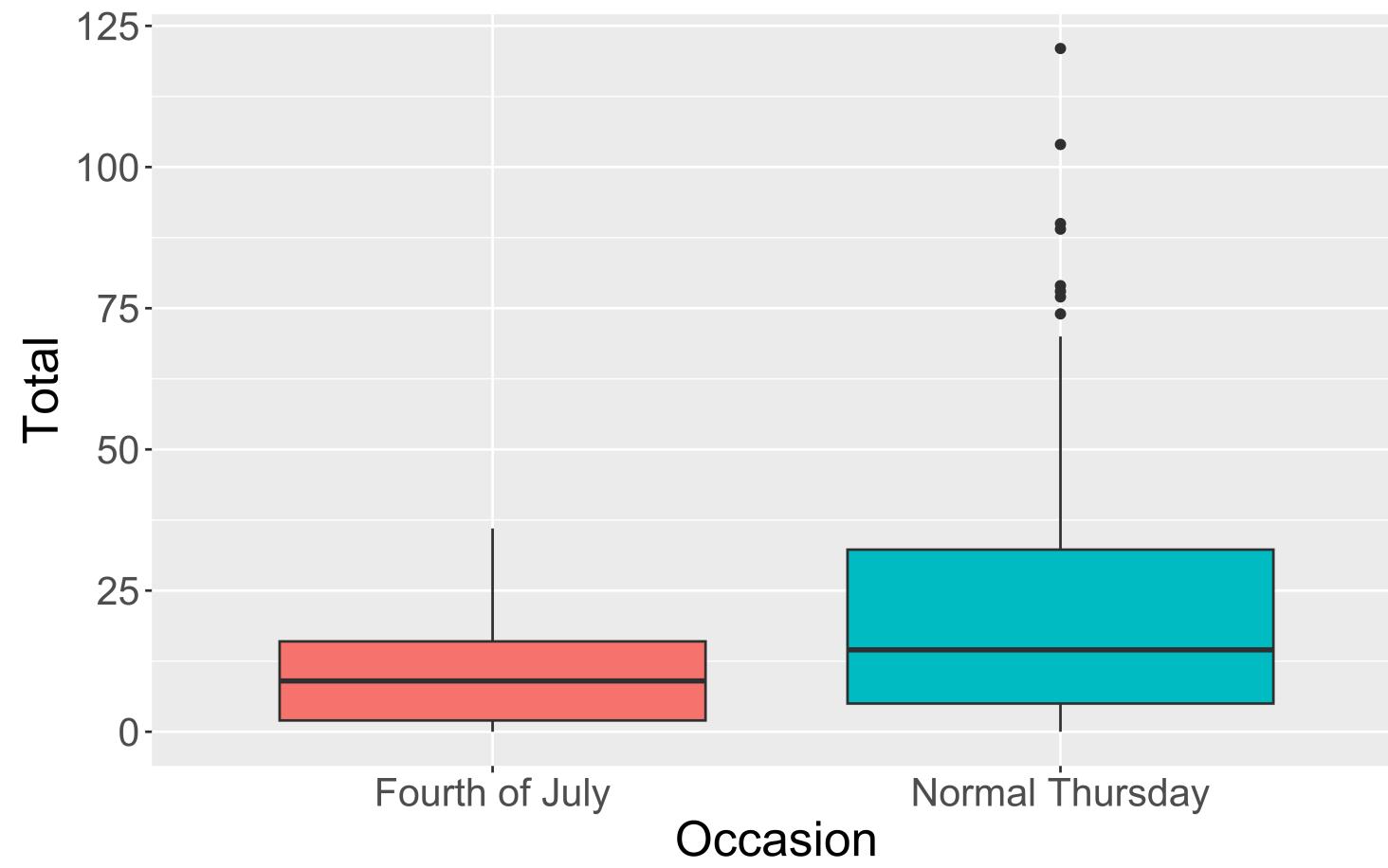


Violin Plots

```
1 ggplot(data = july_2019,  
2         mapping = aes(x = Occasion,  
3                           y = Total,  
4                           fill = Occasion)) +  
5   geom_violin() +  
6   guides(fill = "none")
```



Boxplot Versus Violin Plots



Recap: ggplot2

```
1 library(tidyverse)
2 ggplot(data = ---, mapping = aes(---)) +
3   geom_---(---)
```

Reminders

- Class in full swing:
 - P-set 1 due on Gradescope by noon on Thursday.
 - Visit office hours and then fill out [this form](#)
 - Panel on careers in data science and public health on Thursday at noon in Taylor 210.

