Data Analysis and Visualization in R

Jenn Schilling

June 5, 2022

Introduction to Workshop

Jenn Schilling

she/her/hers

Senior Research Analyst, University Analytics & Institutional Research, University of Arizona

Adjunct Faculty, College for Creative Studies

Schedule

Time	Activity
12:30 - 1:15	Introduction to Workshop and R & Data Analysis in R
1:15 - 1:25	BREAK
1:25 - 2:10	Data Analysis in R & Data Visualization Best Practices
2:10 - 2:20	BREAK
2:20 - 3:05	Data Visualization Best Practices & Data Visualization in R
3:05 - 3:15	BREAK
3:15 - 4:00	Data Visualization in R & Next Steps

air

Learning Outcomes

- Manipulate and analyze data in R
- Explain the grammar of graphics in R
- Create at least three different charts in R
- Develop polished, presentation-ready visualizations in R

Workshop Website

https://bit.ly/air_R

Introductions

https://padlet.com/jschilling_ccs/air_R_intro

Introduction to R and Packages

R

- Why R?

- Packages
  - {here}
  - {explore}
  - {tidyverse}
  - {scales}

here

Illustration by Allison Horst

explore

https://github.com/rolkra/explore

Interactive Exploratory Data Analysis

Automated Report of Data

air

tidyverse

https://www.tidyverse.org/

A collection of R packages for data science

We will be using

- readr
- dplyr
- tidyr
- ggplot2

R Markdown

https://rmarkdown.rstudio.com/

Creates reproducible documents that include code, text, and output all in a single report.

R Setup

- ✓ R
- ✓ RStudio
- ✓ Packages Installed (here, explore, tidyverse, scales)
- ❑ Workshop folder with R Markdown and data files

https://bit.ly/air_R

Workshop Materials

- code
  - workshop-code-along.Rmd
- data
  - IPEDS_inst_adm.csv
  - IPEDS_inst_char.csv
  - IPEDS_inst_compl.csv
  - IPDES_inst_enrl.csv
- output
- plots
- .here
- README.txt

Data

- IPEDS
  - Institutional Characteristics
  - Admissions and Test Scores
  - Fall Enrollment
  - Completions
  - 2017-2020
  - 16 Institutions
  - Reformatted to be tidy

Data Analysis in R

air

Data Exploration

Data Exploration – Your Turn

BREAK

Data Analysis

- Summary Statistics
- Data Manipulation
- Tidy Data

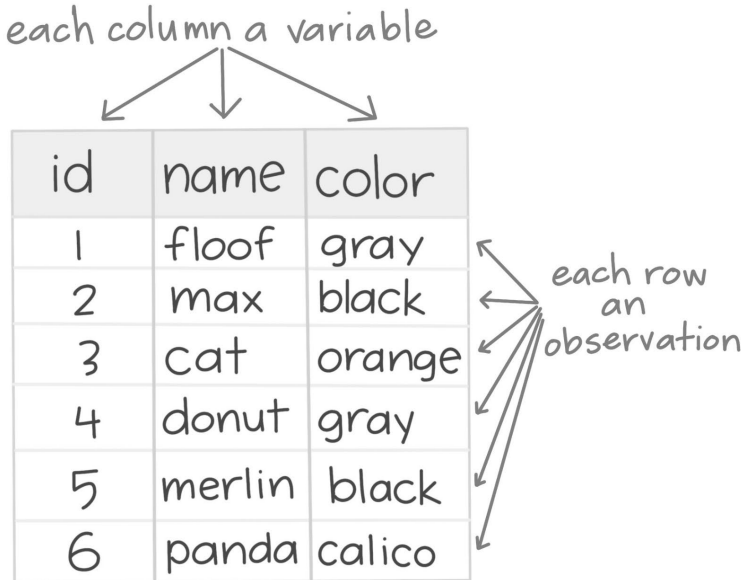
air

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

–HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

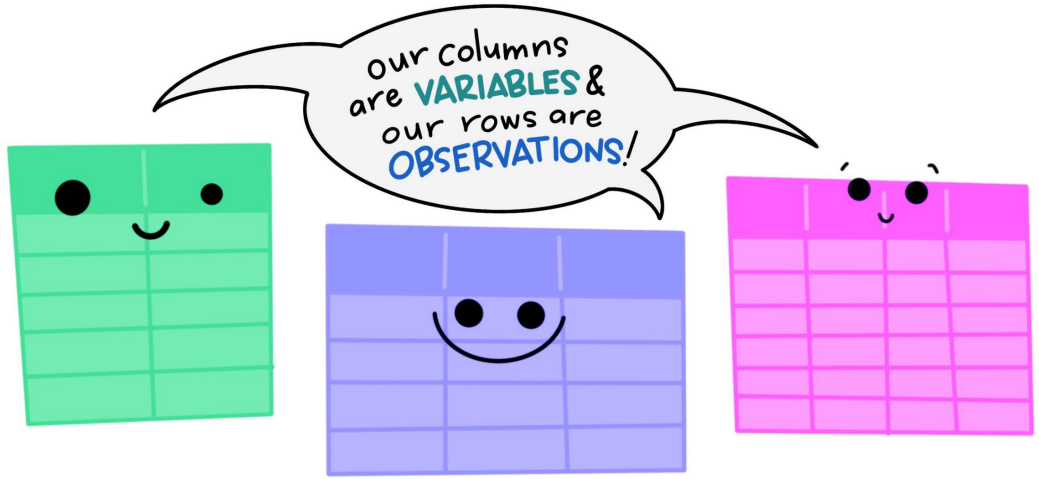


Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

air

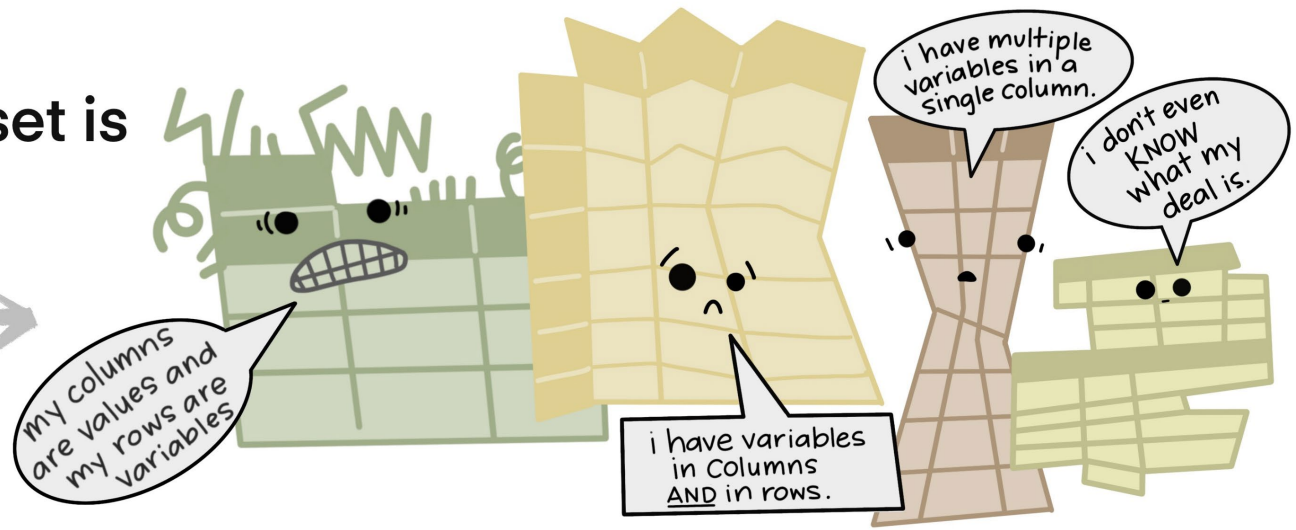
Illustrations from the Openscapes blog *Tidy Data for reproducibility, efficiency, and collaboration* by Julia Lowndes and Allison Horst

The standard structure of tidy data means that "tidy datasets are all alike..."



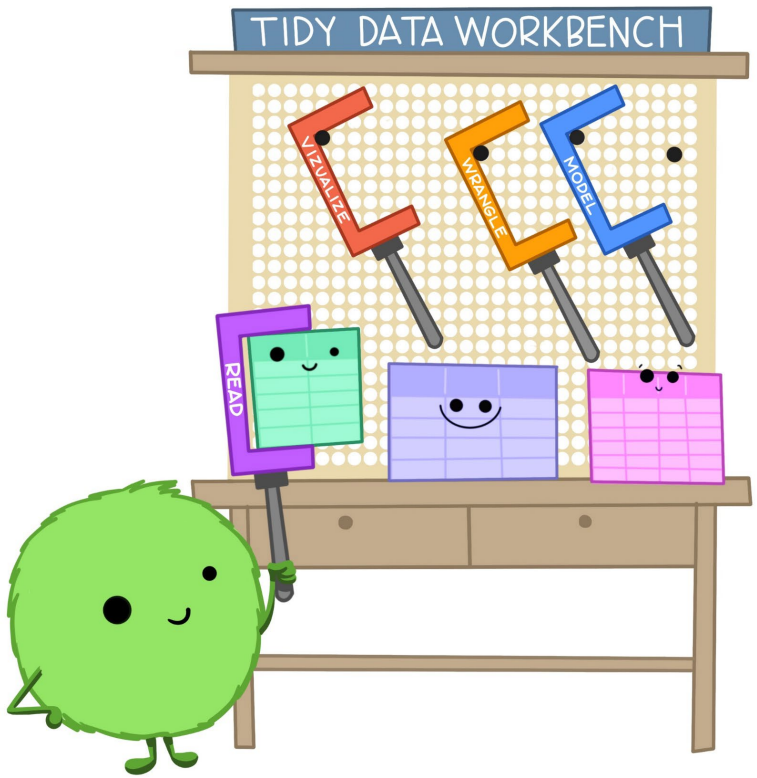
"...but every messy dataset is messy in its own way."

–HADLEY WICKHAM

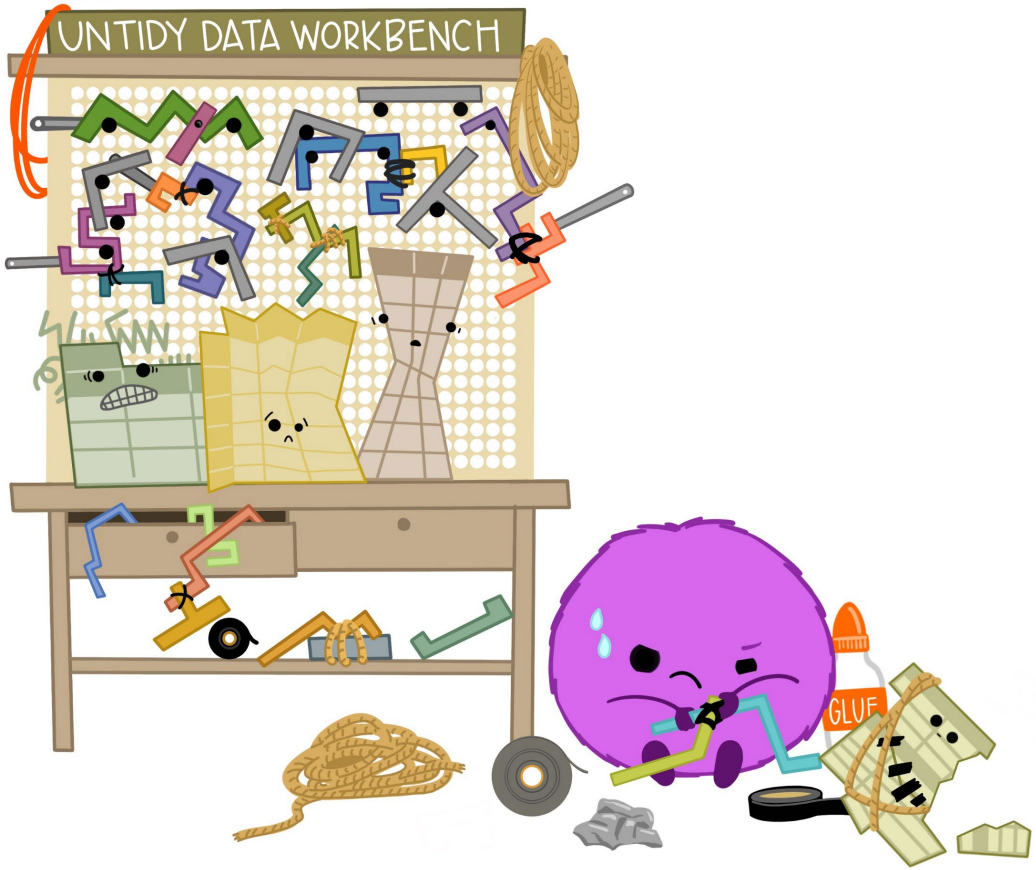


Illustrations from the Openscapes blog *Tidy Data for reproducibility, efficiency, and collaboration* by Julia Lowndes and Allison Horst

When working with tidy data, we can use the same tools in similar ways for different datasets...



...but working with untidy data often means reinventing the wheel with one-time approaches that are hard to iterate or reuse.



Illustrations from the Openscapes blog *Tidy Data for reproducibility, efficiency, and collaboration* by Julia Lowndes and Allison Horst

Data Analysis – Your Turn

Data Visualization Best Practices

https://padlet.com/jaschilling/air_R

What is data visualization?

A Definition from Tableau

“Data visualization is the graphical representation of information and data.” (Tableau, n.d., para. 1)

A Definition from Alberto Cairo

“A data visualization is a display of data designed to enable analysis, exploration, and discovery.” (Cairo, 2016, p. 31)

“A chart is a display in which data are encoded with symbols that have different shapes, colors, or proportions.” (Cairo, 2016, p. 28)

A Definition from W.E.B. Du Bois

“the rendering of information in a visual format to help communicate data while also generating new patterns and knowledge through the act of visualization itself” (Battle-Baptiste & Rusert, 2020, p. 8)

A Definition from Tamara Munzner

“visual representations of datasets intended to help people carry out some task more effectively” (Munzner, 2011, slide 2)

Why visualize data?

Why visualize data?

- Understand the data
- Show patterns
- Engage audience
- Tell a story

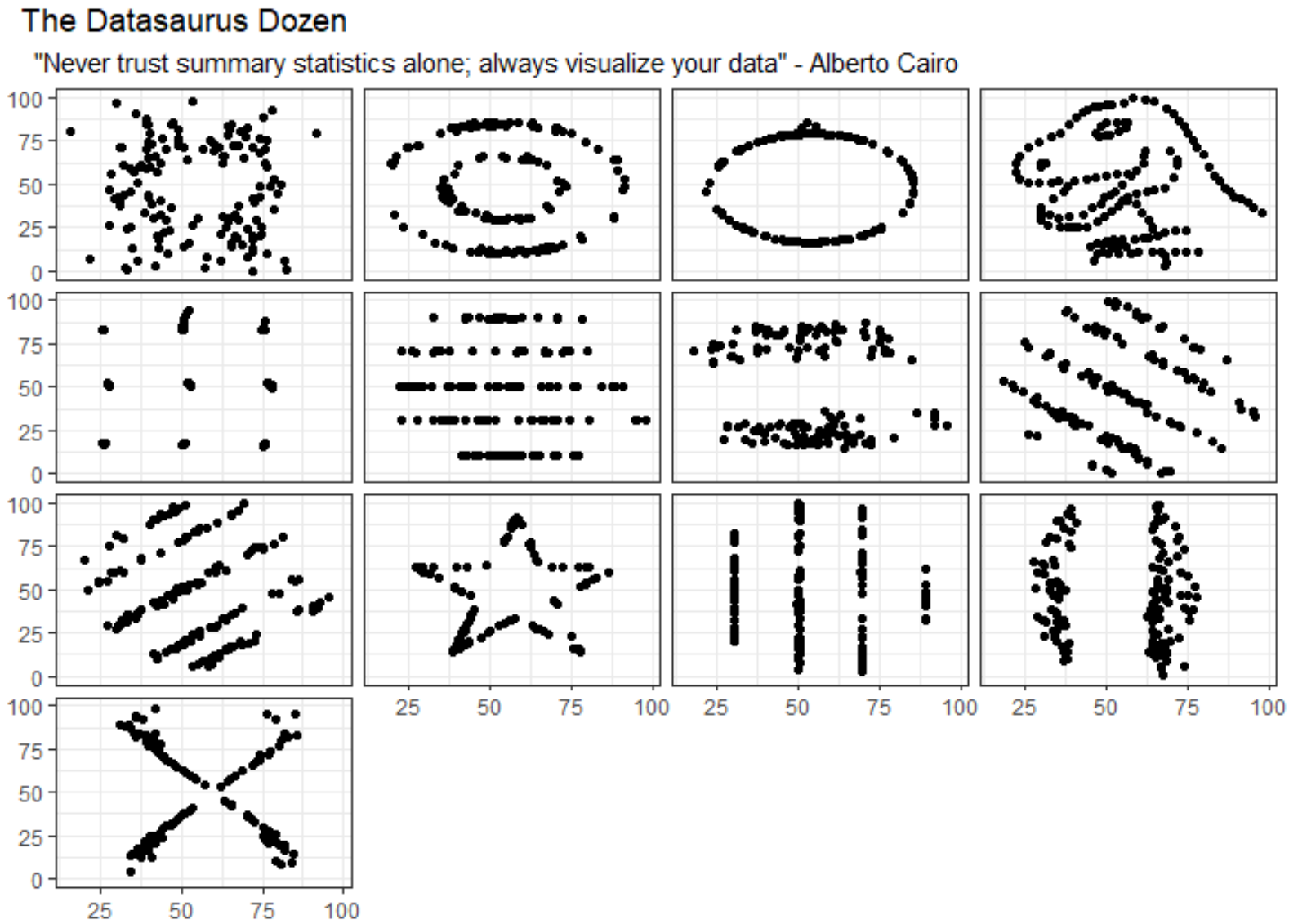
13 datasets of (x, y) coordinate pairs

x	y
55.3846	97.1795
51.5385	96.0256
46.1538	94.4872
42.8205	91.4103
40.7692	88.3333
38.7179	84.8718
35.6410	79.8718
33.0769	77.5641
28.9744	74.4872
26.1538	71.4103

Dataset ID	Mean X	Mean Y	SD X	SD Y	Corr. (x,y)
1	54.27	47.83	16.77	26.94	-0.06
2	54.27	47.83	16.77	26.94	-0.07
3	54.27	47.84	16.76	26.93	-0.07
4	54.26	47.83	16.77	26.94	-0.06
5	54.26	47.84	16.77	26.93	-0.06
6	54.26	47.83	16.77	26.94	-0.06
7	54.27	47.84	16.77	26.94	-0.07
8	54.27	47.84	16.77	26.94	-0.07
9	54.27	47.83	16.77	26.94	-0.07
10	54.27	47.84	16.77	26.93	-0.06
11	54.27	47.84	16.77	26.94	-0.07
12	54.27	47.83	16.77	26.94	-0.07
13	54.26	47.84	16.77	26.93	-0.07

air

Datasets Visualized



What is an effective data visualization?

Effective Data Visualization

“The greatest value of a picture is when it forces us to notice what we never expected to see.”

(Tukey, 1977,p. vi)

Effective Data Visualization

"Ultimately, it is content that makes graphics interesting. When a chart is presented properly, information just flows to the viewer in the clearest and most efficient way. There are no extra layers of color, no enhancements to distract us from the clarity of the information."

(Wong, 2010, p. 13)

Effective Data Visualization

"A good visualization is:

1. reliable information,
2. visually encoded so relevant patterns become noticeable,
3. organized in a way that enables at least some exploration, when it's appropriate,
4. and presented in an attractive manner, but always remembering that honesty, clarity, and depth come first."

(Cairo, 2016, p. 12)

Effective Data Visualization

"The effectiveness of any particular graph is not just a matter of how it looks in the abstract but also a question of who is looking at it, and why."

(Healy, 2019, p. 1)

"A good visualization can do more than just answer questions; it can help you see that there are other questions you need to answer."

(Wexler, 2021, p. 2)

Effective Data Visualization

“Rather than making universal rules and ratios that exclude some aspects of human experience in favor of others, our time is better spent working toward a more holistic and more inclusive ideal. All design fields, including visualization and data communication, are fields of possibility.

Rebalancing emotion and reason opens up the data communication toolbox and allows us to focus on what truly matters in a design process: honoring context, architecting attention, and taking action to defy stereotypes and reimagine the world.”

(D’Iganzio & Klein, 2020, p.96)

BREAK

What do you need to know to visualize data?

- Preattentive Attributes
- Gestalt Principles
- Color Use in Data Visualization
- Types of Data Visualizations
- Hierarchy of Perceptual Tasks

air

How many times does the number 4 appear?

5 4 1 4 0 9 2 5 0 2 0 2

1 1 1 6 8 9 5 6 8 6 7 5

1 1 8 4 2 6 1 1 0 7 3 8

3 5 8 7 2 7 0 2 2 8 3 4

How many times does the number 4 appear?

5 4 1 4 0 9 2 5 0 2 0 2

1 1 1 6 8 9 5 6 8 6 7 5

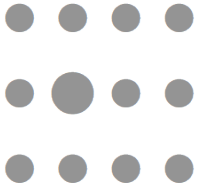
1 1 8 4 2 6 1 1 0 7 3 8

3 5 8 7 2 7 0 2 2 8 3 4

Preattentive Attributes

Length

Width



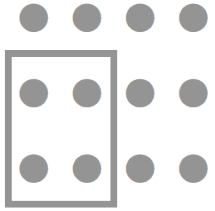
Size

Shape

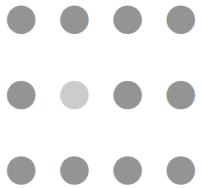
Orientation

Curvature

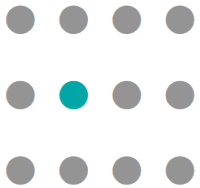
Added Marks



Enclosure



Intensity



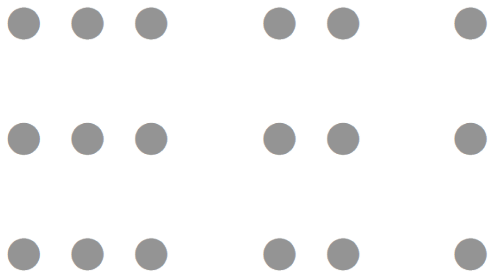
Hue

Position

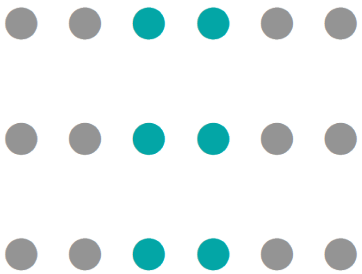
Adapted from Stephen Few's Tapping the Power of Visual Perception (2004).

air

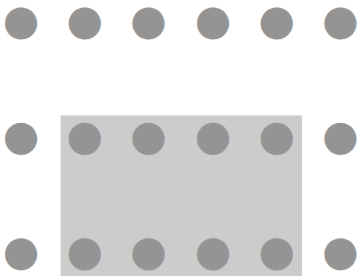
Gestalt Principles



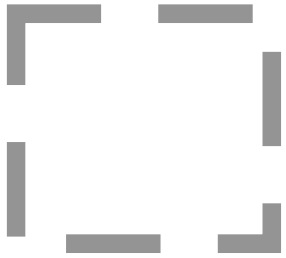
Proximity



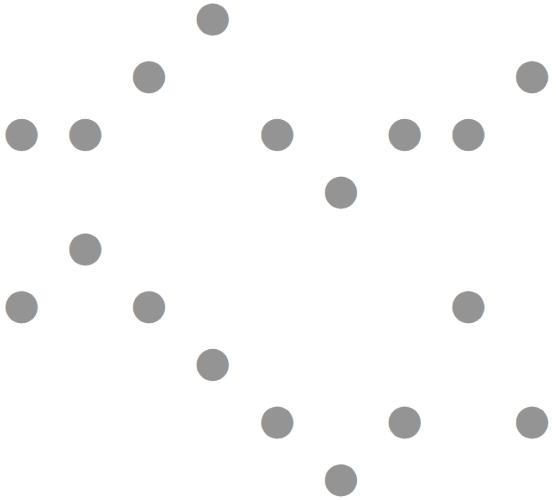
Similarity



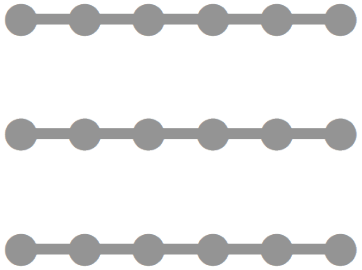
Enclosure



Closure



Continuity

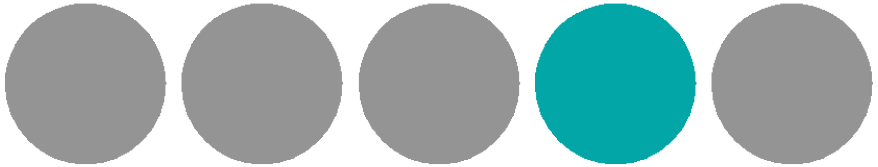


Connection

air

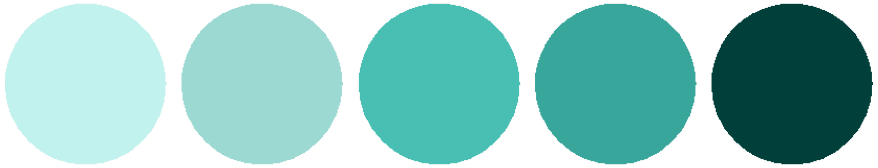
Color Use in Data Visualization

Highlight



Categorical

Sequential



Diverging

Adapted from Steve Wexler's The Big Picture (2021)

air

How do you know which chart to use?

"When it comes to selecting a graph, first and foremost, choose a graph type that will enable you to clearly get your message across to your audience." (Knaflic, 2015, p. 60)

"The purpose of your graphics should somehow guide your decision of how to shape the information." (Cairo, 2016, p. 51)

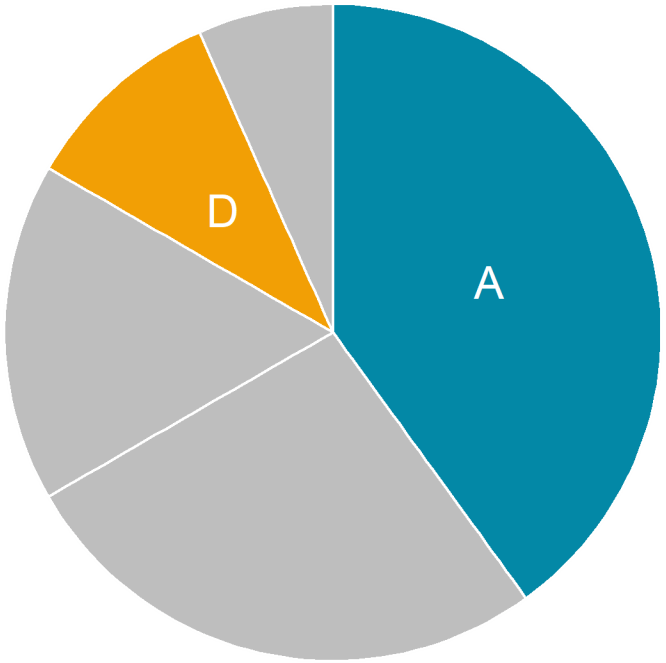
Resources

- From Data to Viz
- Depict Data Studio – Chart Chooser
- Visual Vocabulary

Perceptual Tasks

Perceptual Tasks

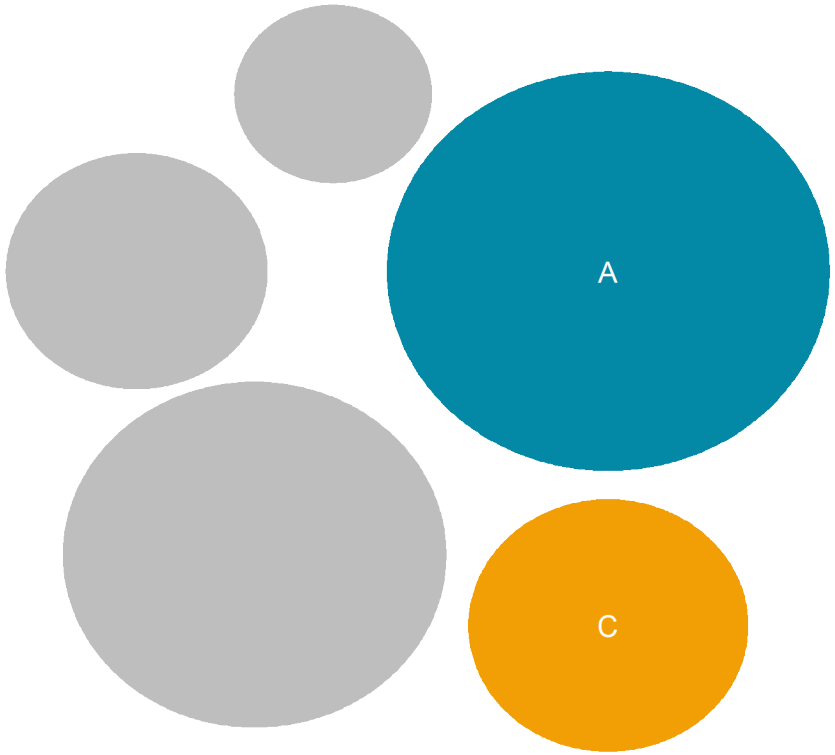
If the value of A is 40%, what is the value of D?



Adapted from Wexler (2021)

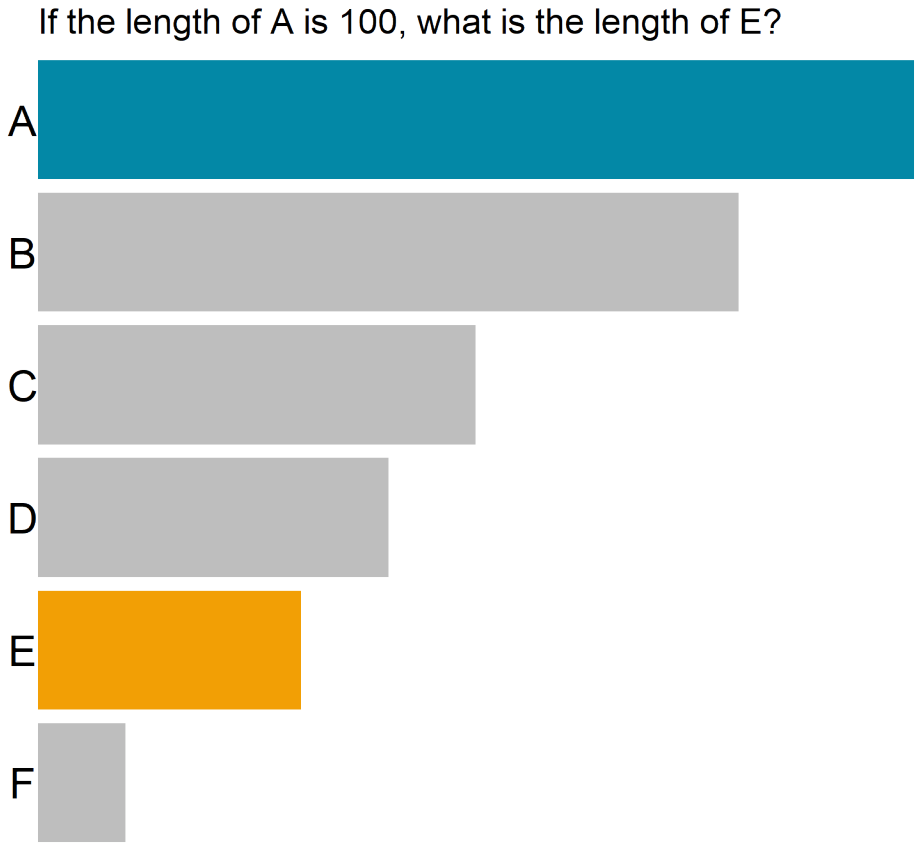
Perceptual Tasks

If the area of A is 100, what is the area of C?



Adapted from Wexler (2021)

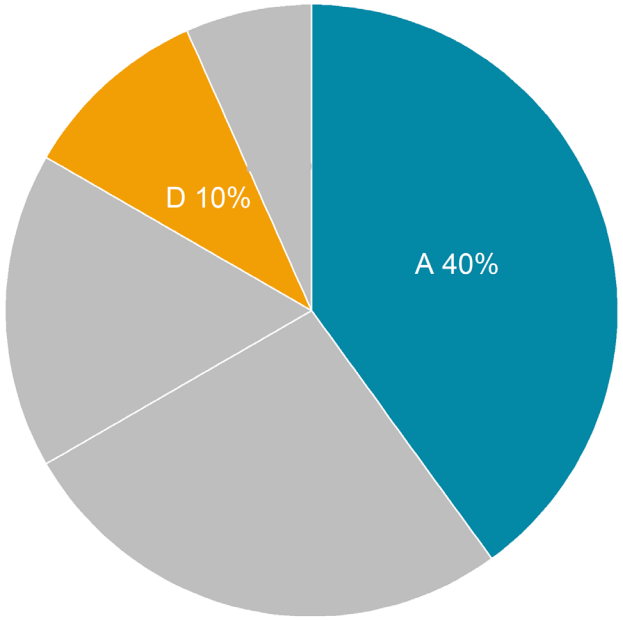
Perceptual Tasks



Adapted from Wexler (2021)

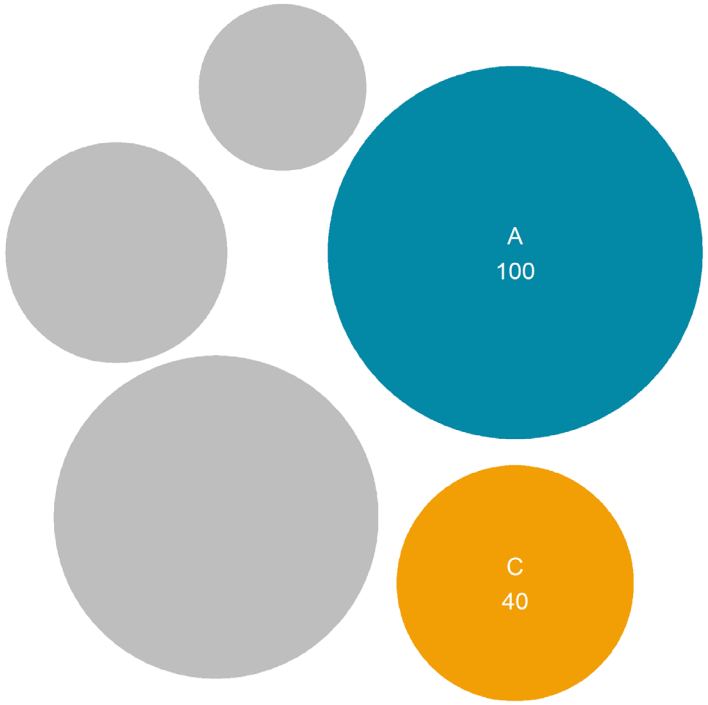
Perceptual Tasks

If the value of A is 40%, what is the value of D?



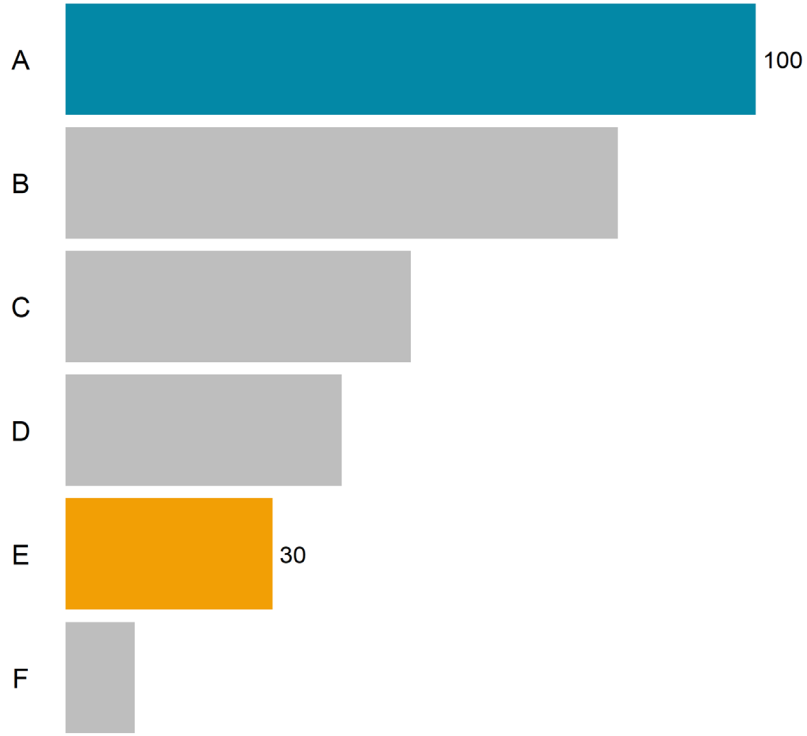
Adapted from Wexler (2021)

If the area of A is 100, what is the area of C?



Adapted from Wexler (2021)

If the length of A is 100, what is the length of E?



Adapted from Wexler (2021)

Perceptual Tasks

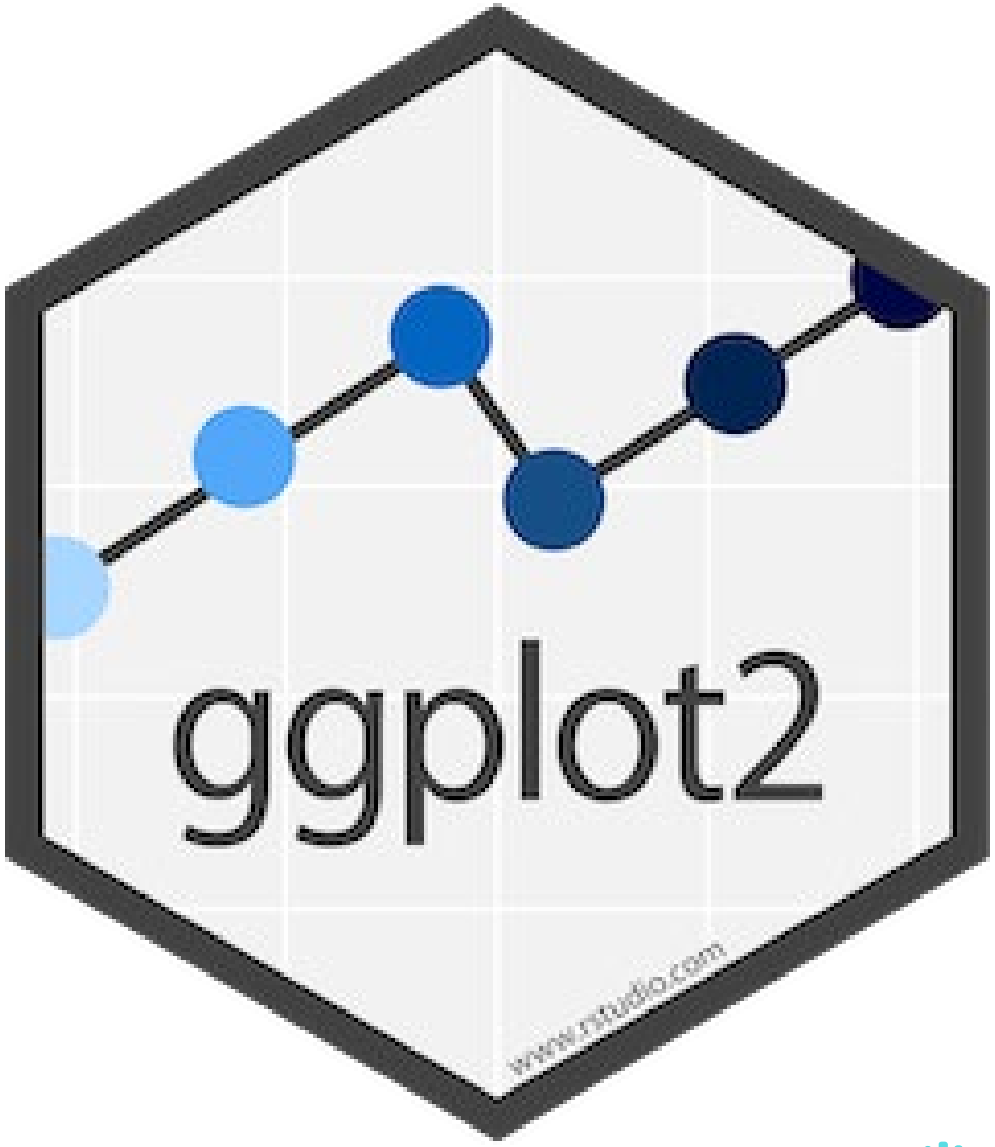
Adapted from Heer & Bostock (2010)

Data Visualization in R

Data Visualization in R

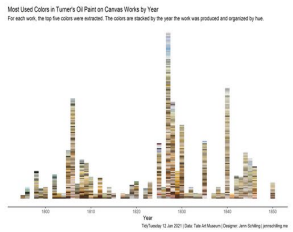
Library: ggplot2

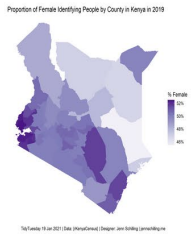
- Robust
- Iterative
- Flexible
- Functional
- Reproducible

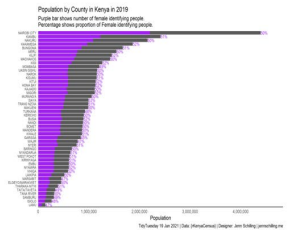


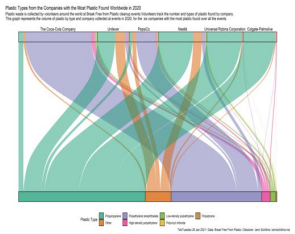
air

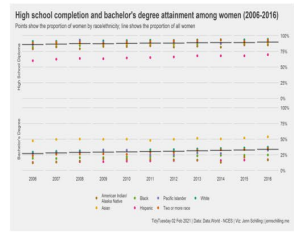
What's Possible with ggplot2?

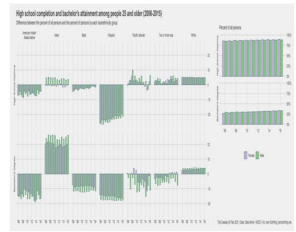


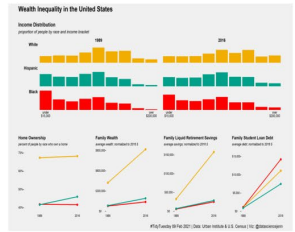


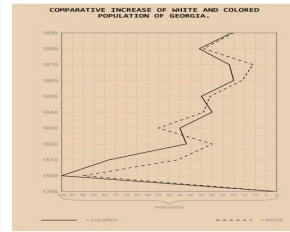


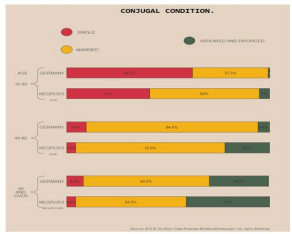


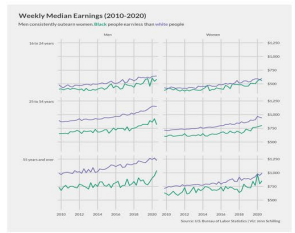


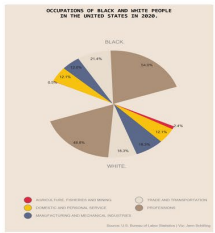


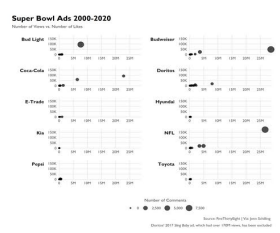


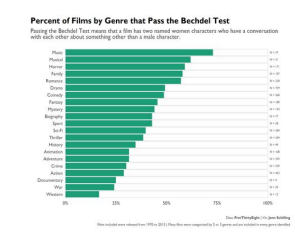


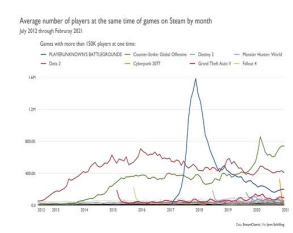


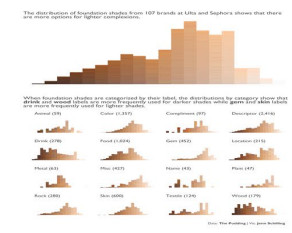


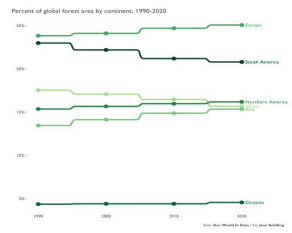


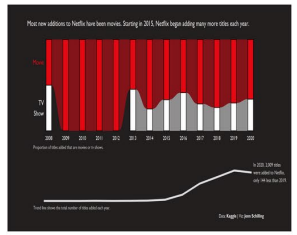


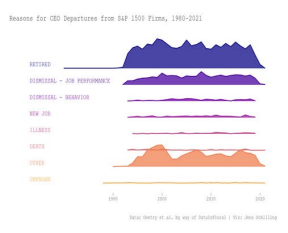


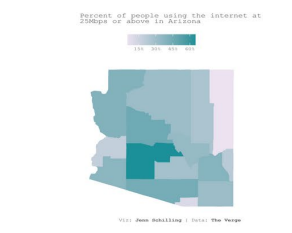


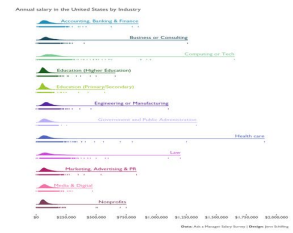


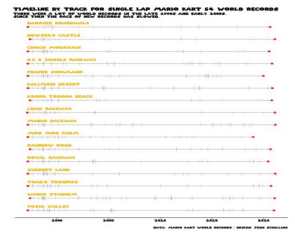


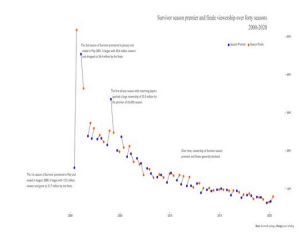


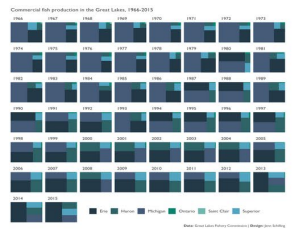


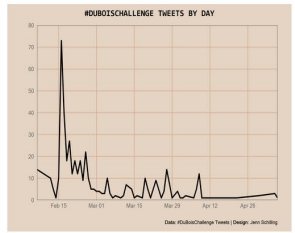


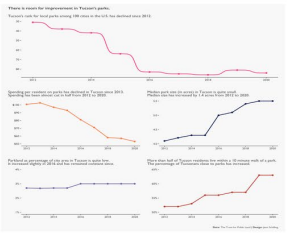


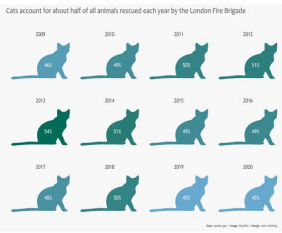


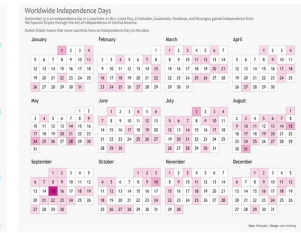


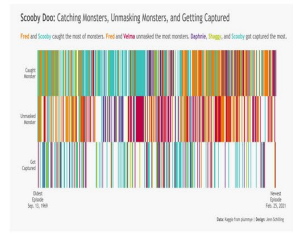


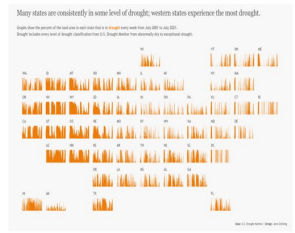


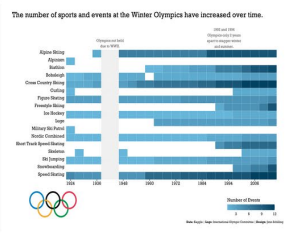


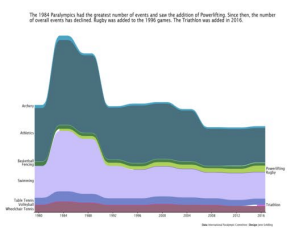


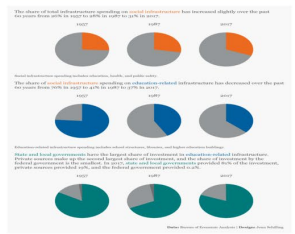


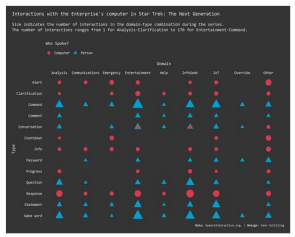


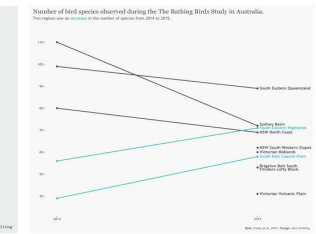


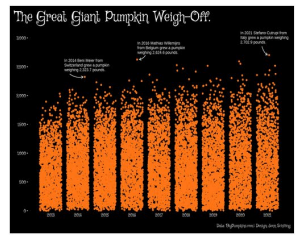


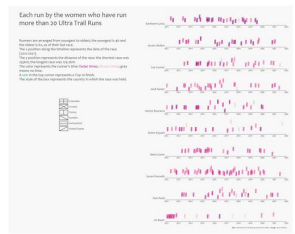


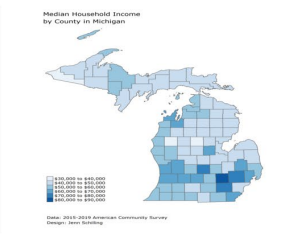




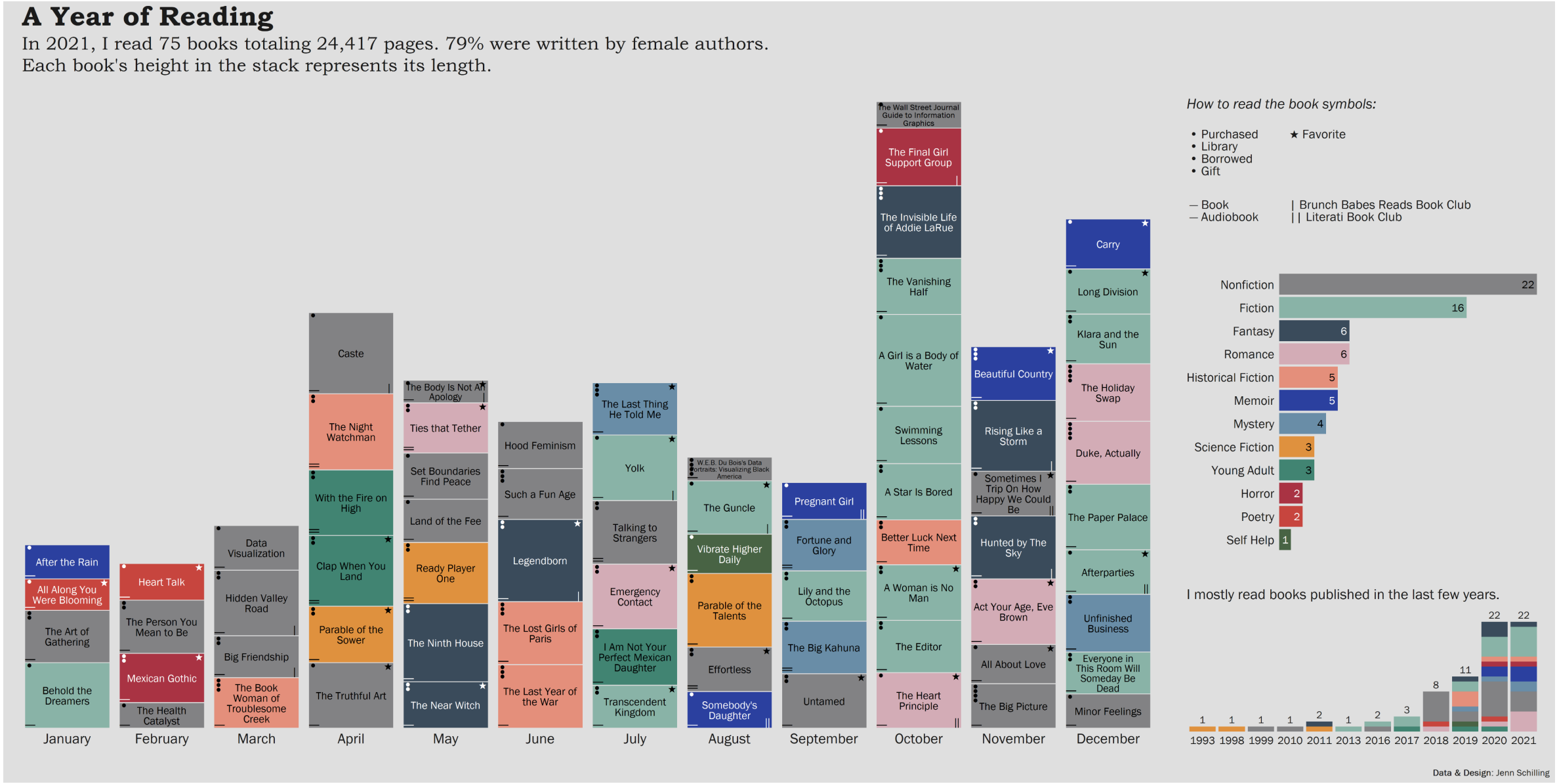








What's Possible with ggplot2?



What's Possible with ggplot2?

November is National Novel Writing Month. Hundreds of thousands of people around the world participate. The goal is to write 50,000 words during the month. In 2021, I participated for the second time. I wrote every day during November. I ended the month with a total of 50,179 words after 33 hours and 41 minutes of writing.

1,672 average words per day.

On the two days I wrote the most words, I spent over 90 minutes writing. These days occurred on weeks that were less balanced; I wrote fewer words on some days and more on others.

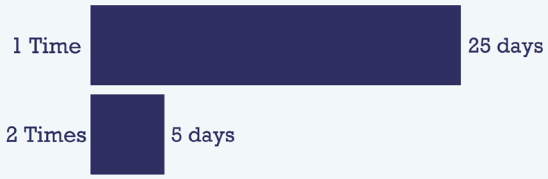
67 average minutes writing per day.

The words came more easily, and I spent less time writing during the third week and the last few days of the month. I struggled more at the start of the month and in the fourth week when it took me longer to meet my daily word count goal.

26 average words per minute each day.

I wrote the fastest in the middle of the month and end of the month. Beginning the story and leading up to the end were my slowest writing times.

I mostly wrote once per day.



I wrote in the morning most frequently.

Logo: Image courtesy of National Novel Writing Month | **Data & Design:** Jenn Schilling | Learn more at nanowrimo.org

The Grammar of Graphics

- The basis of ggplot2
- Build plots in layers:
  - Data to be visualized
  - Aesthetic mappings
  - Geometric objects
  - Statistical transformations
  - Coordinates
  - Scales
  - Facets

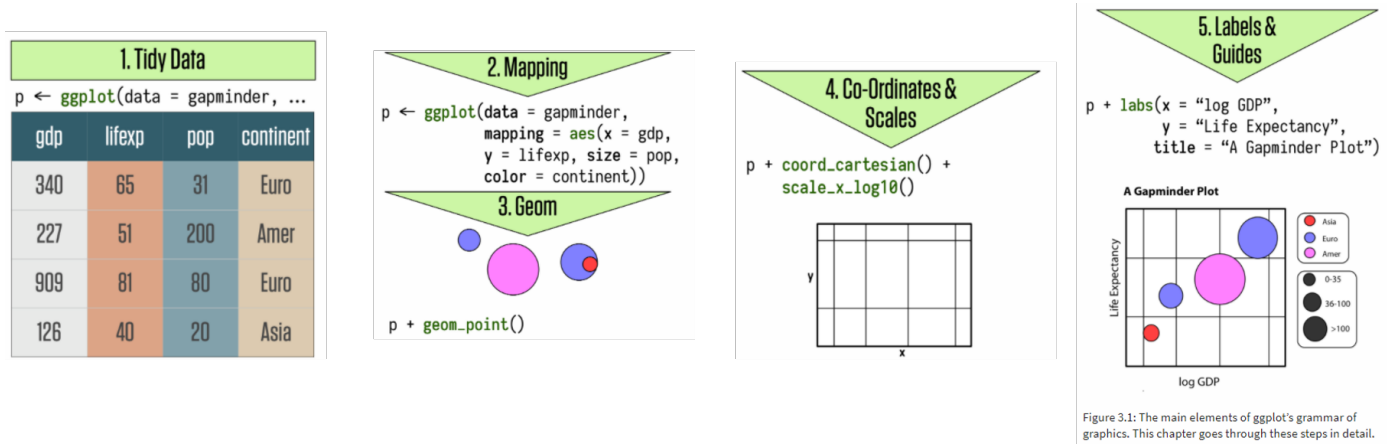
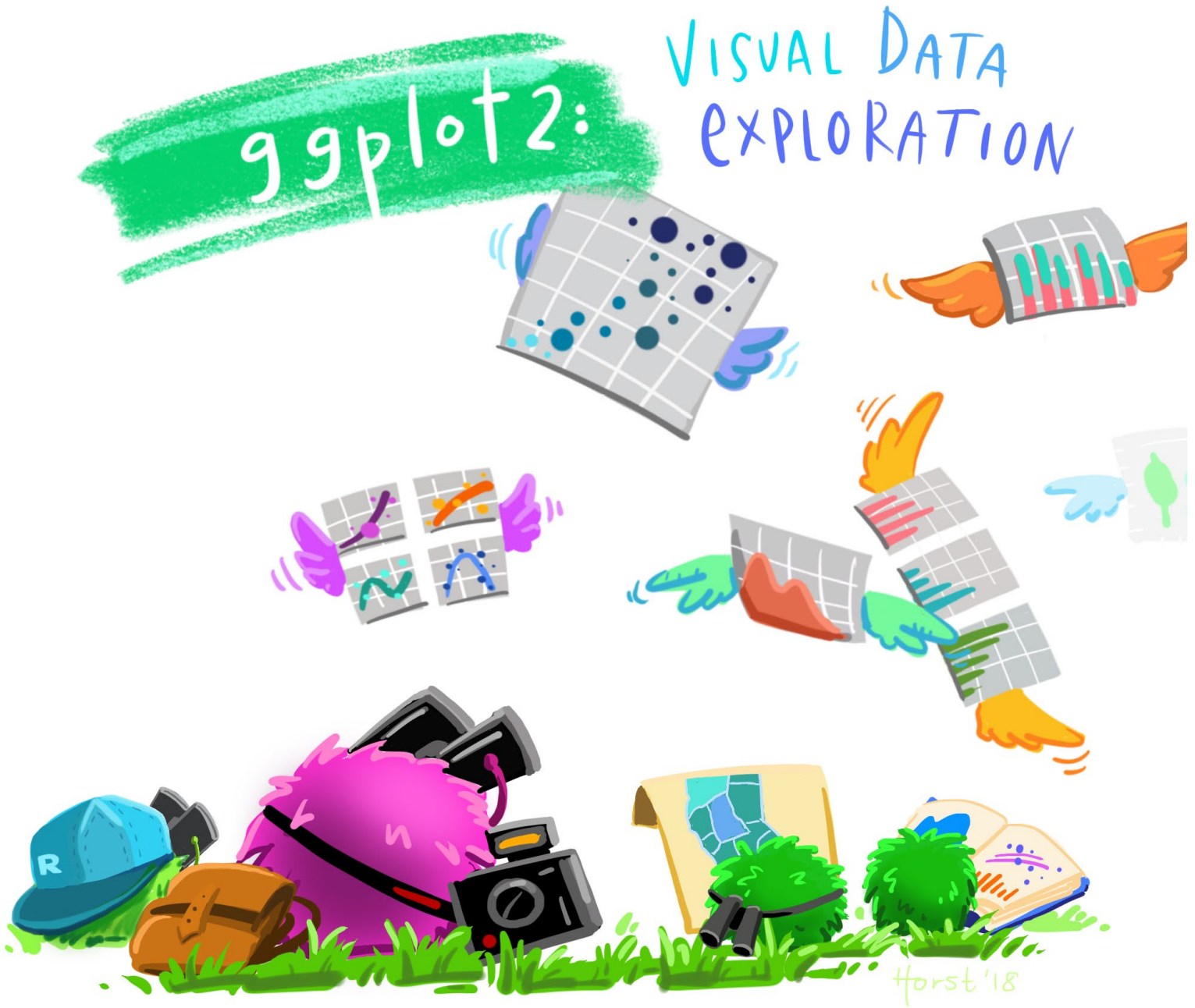


Figure 3.1: The main elements of ggplot's grammar of graphics. This chapter goes through these steps in detail.

(Wickham, 2010, p. 8; Healy, 2019, p. 56)



Artwork by @allison_horst

air

Artwork by @allison_horst

Data Visualization

BREAK

The Grammar of Graphics

- The basis of ggplot2
- Build plots in layers:
  - Data to be visualized
  - Aesthetic mappings
  - Geometric objects
  - Statistical transformations
  - Coordinates
  - Scales
  - Facets

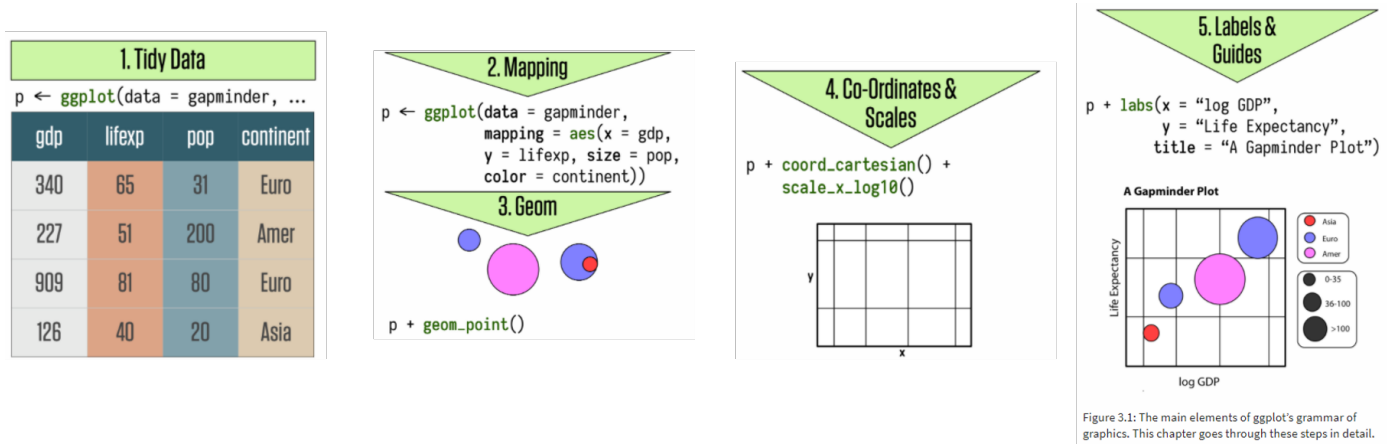


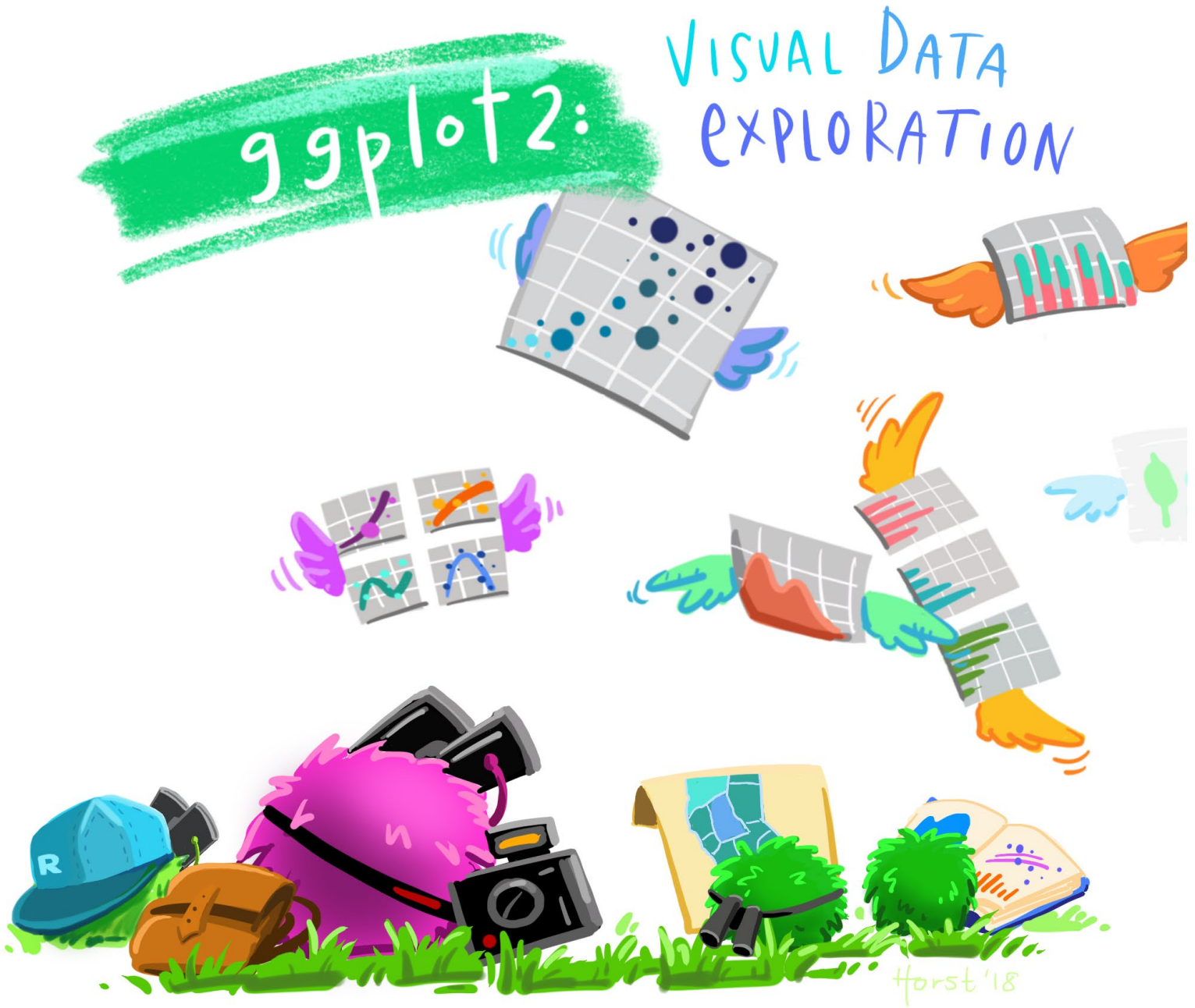
Figure 3.1: The main elements of ggplot's grammar of graphics. This chapter goes through these steps in detail.

(Wickham, 2010, p. 8; Healy, 2019, p. 56)

Data Visualization

Data Visualization – Your Turn

Artwork by @allison_horst



Artwork by @allison_horst

air

Where to go next

Experiment with the provided code and data sources. There are many other plots that could be made using this data.

Some Resources:

- ggplot2: Elegant Graphics for Data Analysis by Hadley Wickham
- Data Visualization Chapter in R for Data Science by Hadley Wickham & Garrett Grolemund
- ggplot2 Reference Guide

What did we learn today?

- Manipulate and analyze data in R
- Explain the grammar of graphics in R
- Create at least three different charts in R
- Develop polished, presentation-ready visualizations in R

Thank you!

jaschilling@arizona.edu
@datasciencejenn

References

Battle-Baptiste, W., & Rusert, B. (2018). *W.E.B. Du Bois's Data Portraits: Visualizing Black America: The color line at the turn of the Twentieth Century.* Princeton Architectural Press.

Cairo, A. (2016). *The truthful art: Data, charts, and maps for communication.* New Riders

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism.* The MIT Press.

Few, S. (2006). *Information dashboard design.* O'Reilly Media, Inc.

Healy, K. (2019). *Data visualization: a practical introduction*. Princeton University Press.

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 203–212). New York, NY, USA: ACM. https://doi.org/10.1145/1753326.1753357

Knaflic, C. N. (2015). *Storytelling with data: A Data Visualization Guide for Business Professionals.* John Wiley & Sons, Inc.

Munzner, T. (2011). *Visualization Principles* [PDF slides]. Department of Computer Science, University of British Columbia. https://www.cs.ubc.ca/~tmm/talks/vizbi11/vizbi11.pdf

Tableau. (n.d.). *What is data visualization? definition, examples, and learning resources.* Tableau. Retrieved January 18, 2022, from https://www.tableau.com/learn/articles/data-visualization

Tufte, E. R. (2001). *The visual display of quantitative information*. Graphics Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Publishing Company.

Wexler, S. (2021). *The big picture: How to Use Data Visualization to Make Better Decisions*. McGraw Hill.

Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics, 19*(1), 3-28. Retrieved March 7, 2021, from http://vita.had.co.nz/papers/layered-grammar.pdf

Wong, D. M. (2013). *The wall street journal guide to information graphics: The dos and don'ts of presenting data, facts, and figures.* Norton.

air