DATA MATTERS™

DATA SCIENCE SHORT COURSE SERIES

AUGUST 7 - 11, 2023

VIRTUAL VIA ZOOM

Welcome!
We'll be
starting soon!

UNC | ODUM INSTITUTE FOR RESEARCH IN SOCIAL SCIENCE

ncds | THE NATIONAL CONSORTIUM for DATA SCIENCE

renci

Visualization for Data Science in R

Angela Zoss

Data Matters Fall 2023

https://www.angelazoss.com/RVis-2Day/

Try right now:
Open RStudio
Try running “library(tidyverse)”
Tell me about any errors

Slides and files

https://github.com/amzoss/RVis-2Day

Schedule, Day 1

Session	Topics	Duration
Session 1	Visualization and data science Intro, setup, basic ggplot2 syntax	9:30 a.m. – 10:35 a.m.
Morning break		10:35 a.m. – 10:50 a.m.
Session 2	Trying more charts	10:50 a.m. – 11:55 a.m.
Lunch		11:55 a.m. – 1:10 p.m.
Session 3	Customizing plots, saving charts out	1:10 p.m. – 2:15 p.m.
Afternoon break		2:15 p.m. – 2:30 p.m.
Session 4	Plot inheritance, advanced examples	2:30 p.m. – 3:35 p.m.
Q&A		3:35 p.m. – 3:40 p.m.

Schedule, Day 2

Session	Topics	Duration
Session 1	ggplot2 review, advanced techniques	9:30 a.m. – 10:35 a.m.
Morning break		10:35 a.m. – 10:50 a.m.
Session 2	Working with text variables	10:50 a.m. – 11:55 a.m.
Lunch		11:55 a.m. – 1:10 p.m.
Session 3	Simple interactive plots	1:10 p.m. – 2:15 p.m.
Afternoon break		2:15 p.m. – 2:30 p.m.
Session 4	Building visualizations into layouts	2:30 p.m. – 3:35 p.m.
Q&A		3:35 p.m. – 3:40 p.m.

Other course logistics

- We all have different skill levels here. That's great!
- Questions and interruptions are welcome, especially if you are lost.
  I want everyone to be able to follow along.
- You may know the answer to someone else's question. If it's quick, feel free to make suggestions in the Zoom chat. Otherwise, I'll be happy to address it myself.
- You may have more advanced suggestions on top of what I'm teaching. Please try not to share these in chat. Too much chat can be a distraction, and I have a specific sequence I follow to keep the content approachable. You can share advanced things in Slack instead.

Set up environment

- R
- RStudio
- packages

Packages:

- tidyverse
- readxl
- markdown
- knitr
- plotly
- colorspace
- dt
- crosstalk
- flexdashboard
- here

Visualization for Data Science

Why visualize in R?

- Quickly explore data
- Save time switching to another tool
- Use charts to inspire new analyses and vice versa
- Reproducibility

Why care about reproducibility?

- Open science makes review easier
- Increasingly a requirement
- Saves you a lot of time trying to figure out what you did last time!

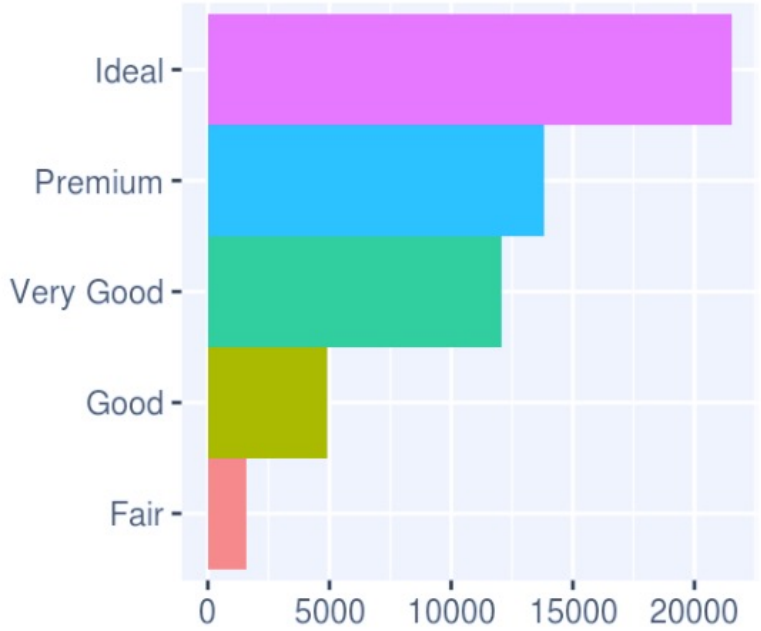
*“Your closest collaborator is **you** six months ago, but you don’t reply to emails.”*

- Mark Holder

ggplot2

What is ggplot2?

an R package designed to create plots based on a theory of the grammar of graphics.



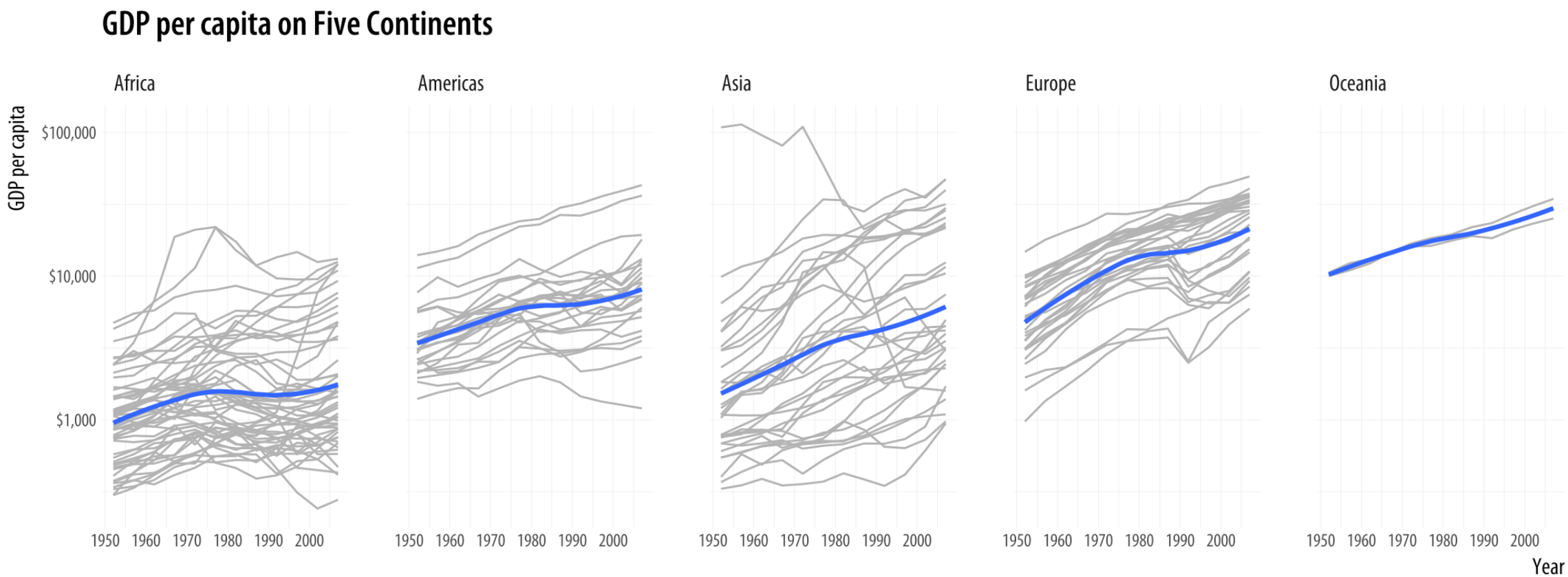
http://r4ds.had.co.nz/data-visualisation.html

Grammar of graphics

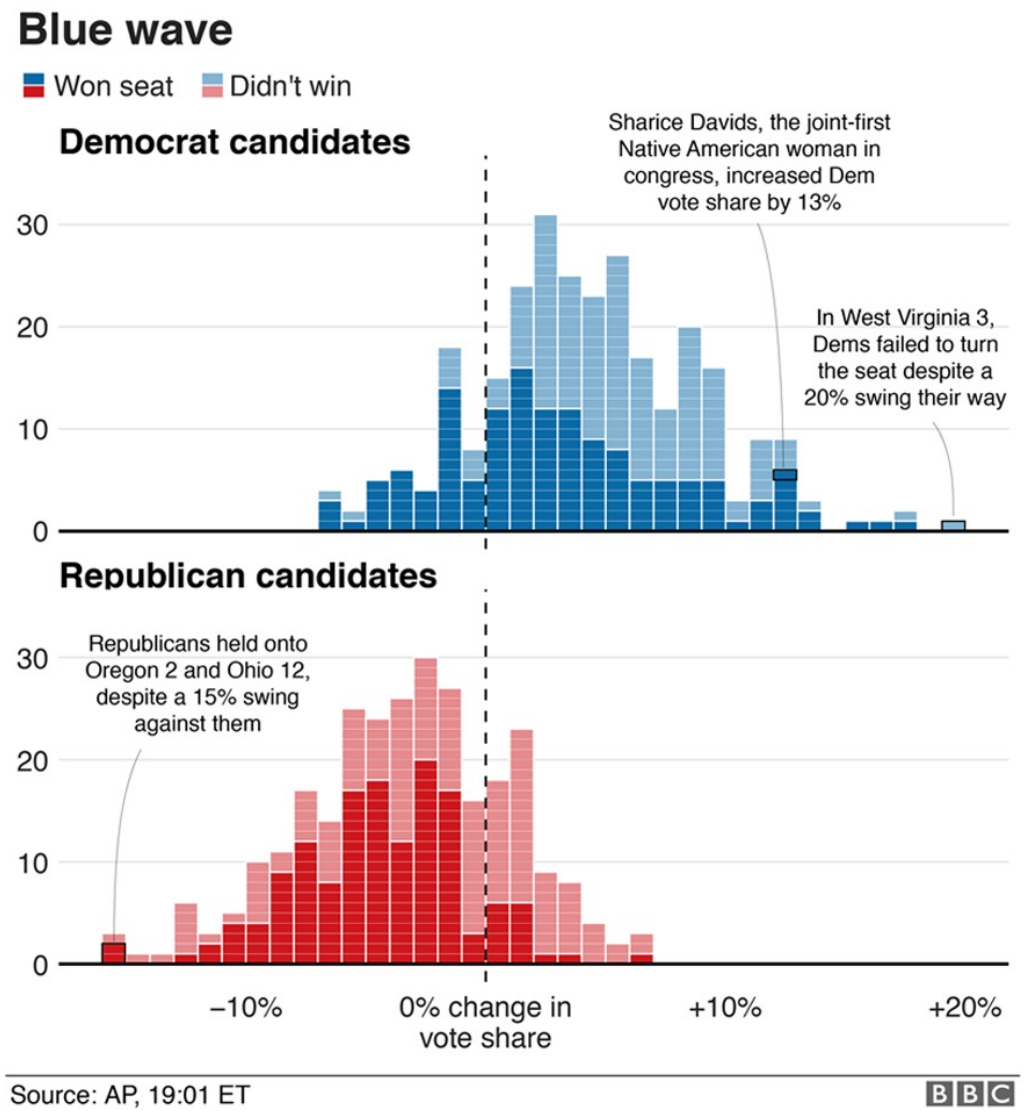
1. DATA: a set of data operations that create variables from datasets
2. TRANS: variable transformations (e.g., rank)
3. SCALE: scale transformations (e.g., log)
4. COORD: a coordinate system (e.g., polar)
5. ELEMENT: graphs (e.g., points) and their aesthetic attributes (e.g., color)
6. GUIDE: one or more guides (axes, legends, etc.).

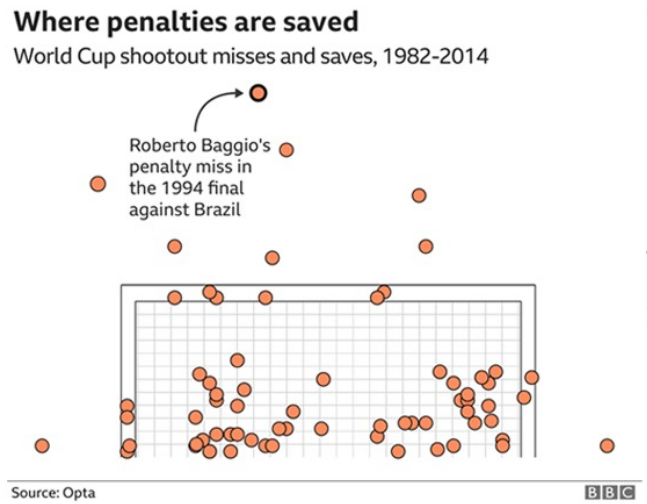
Wilkinson, Leland. (2005). *The grammar of graphics (2nd ed)*. New York: Springer.

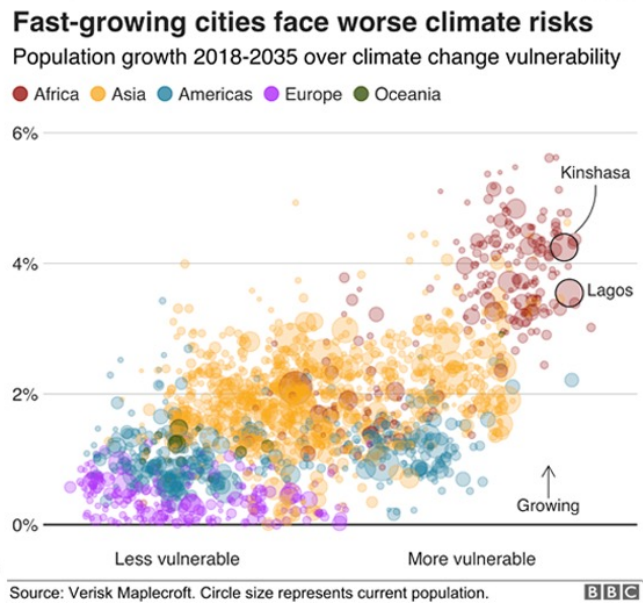
ggplot2 examples

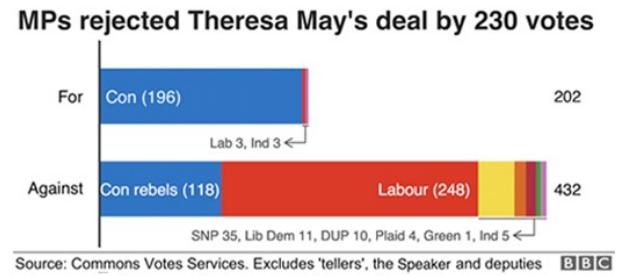


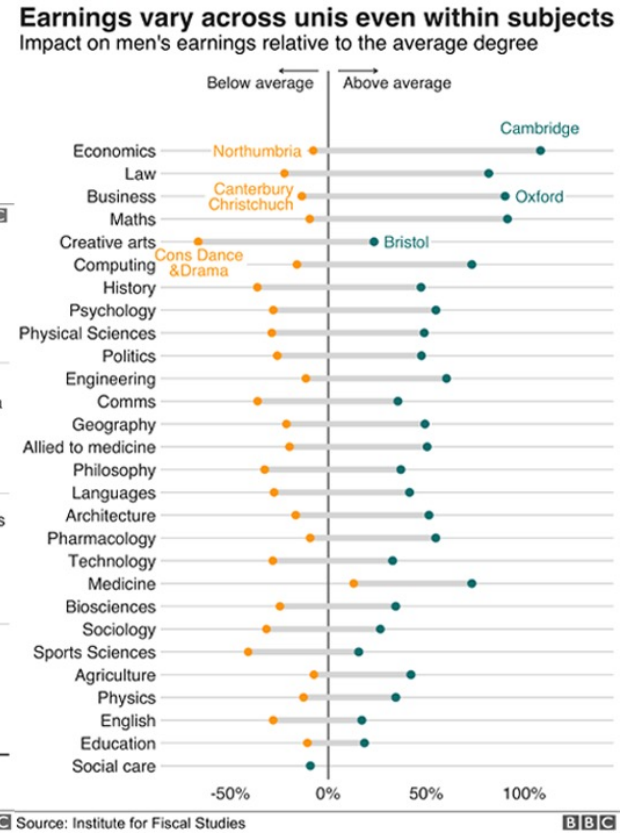
http://socviz.co/groupfacettx.html





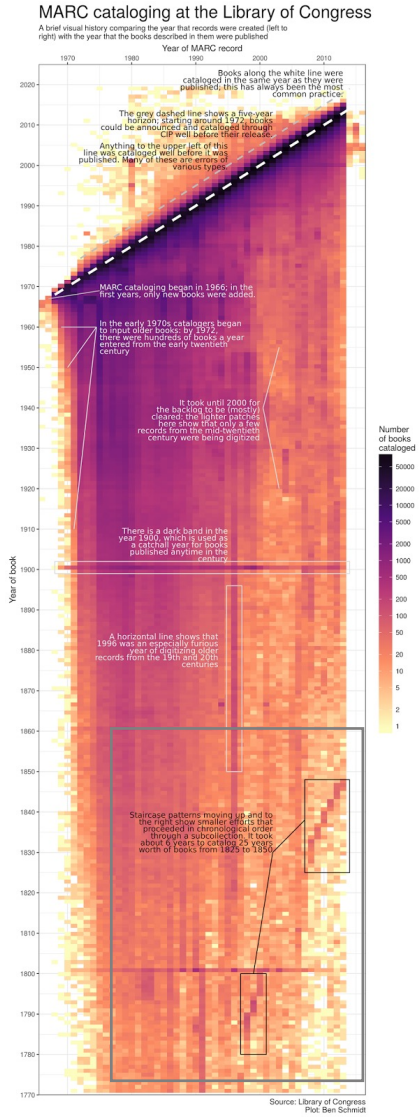


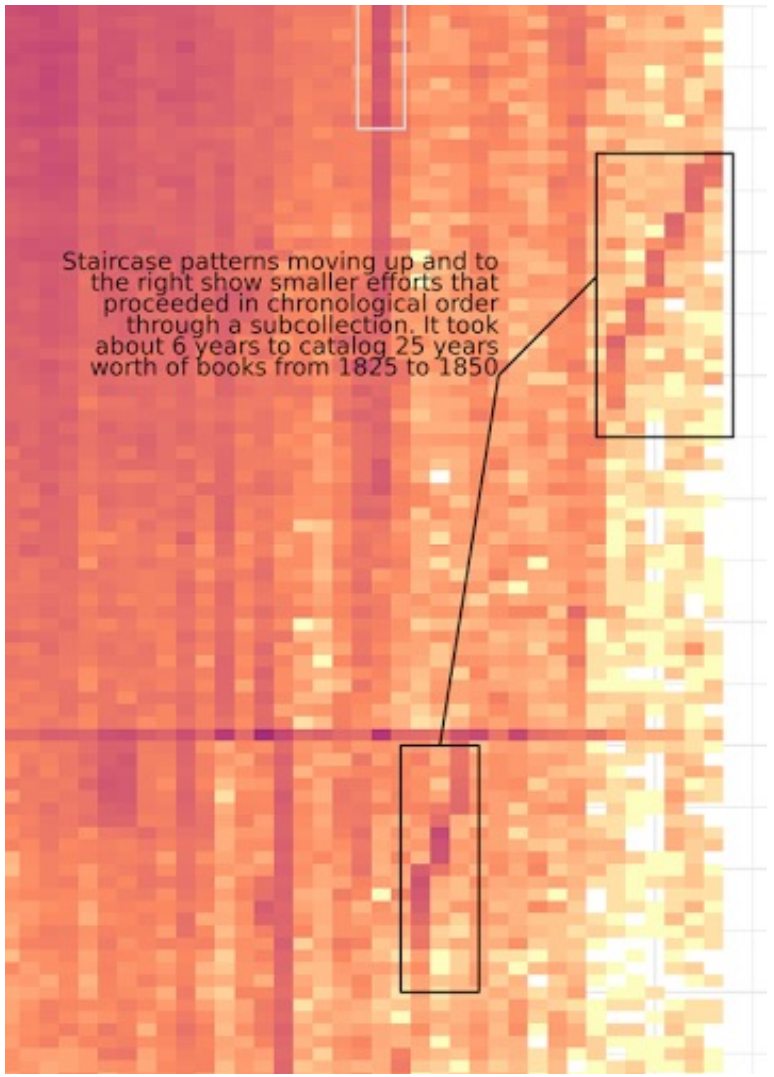




https://bbc.github.io/rcookbook/

https://web.archive.org/web/20150702141747/http://spatial.ly/2015/03/mapping-flows/





http://sappingattention.blogspot.com/2017/05/a-brief-visual-history-of-marc.html

Why ggplot2 instead of base R?

- nice defaults
- easy faceting
- (arguably) more natural syntax
- can switch chart types more easily

“Why I use ggplot2”, David Robinson
http://varianceexplained.org/r/why-I-use-ggplot2/

R vs. Excel, Tableau, etc.

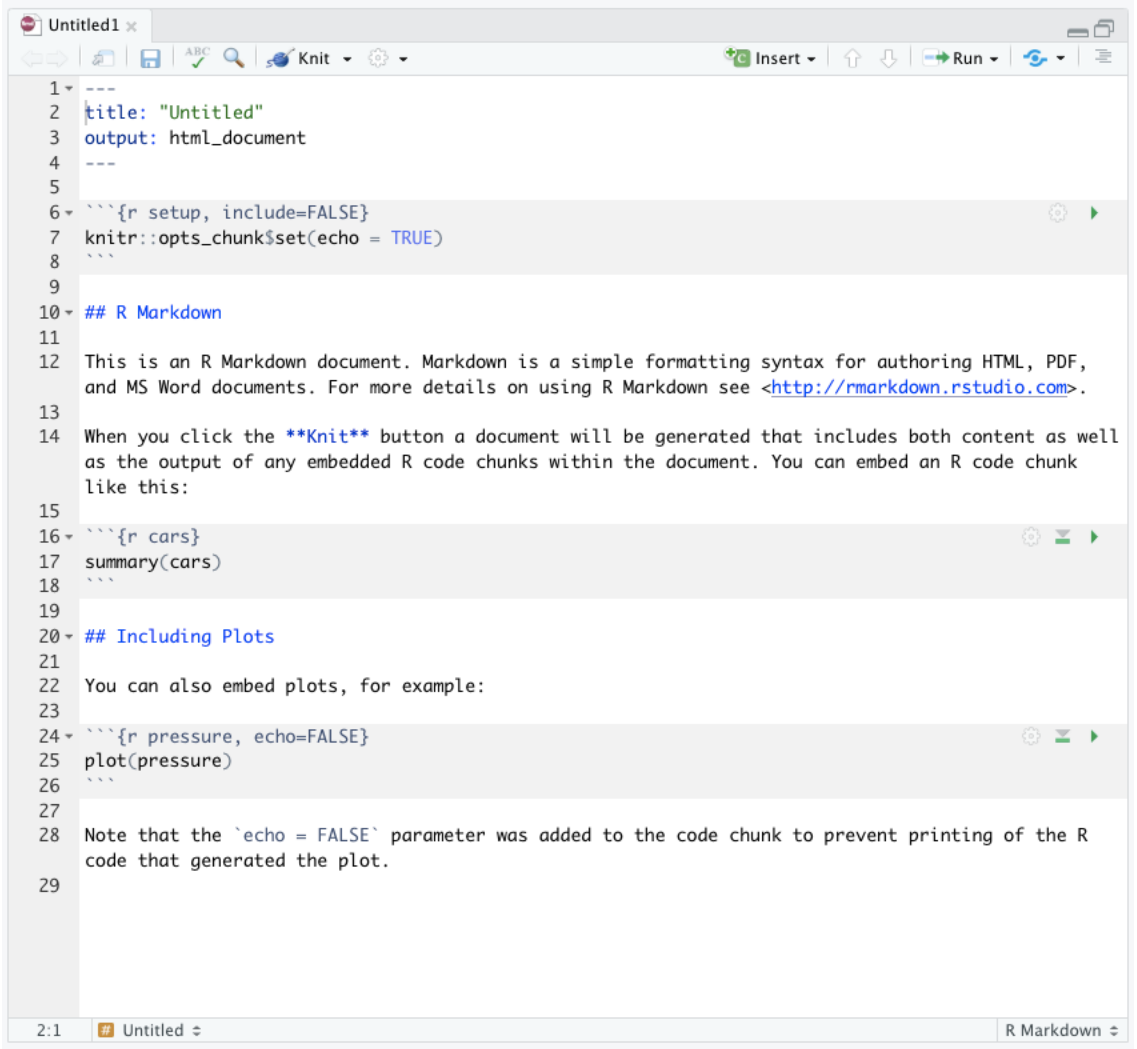
Questions to ask:

- Are you already using R? Why switch?
- Are you going to have to share this process or reproduce it? Try R!
- Is it a quick project, or will others work on it? Maybe Excel is fine.
- Do you need to try a bunch of charts quickly, build interactive components, etc.? Tableau might be more powerful and faster.

Working in RStudio

R Markdown files

- Blend “normal” text (using Markdown syntax for formatting) with code chunks and their output
- Can be compiled (“knit”) into other formats (HTML, Word, PDF)
- Similar to Jupyter Notebooks for Python
- NB: The next generation of R Markdown is Quarto



Why R Markdown?

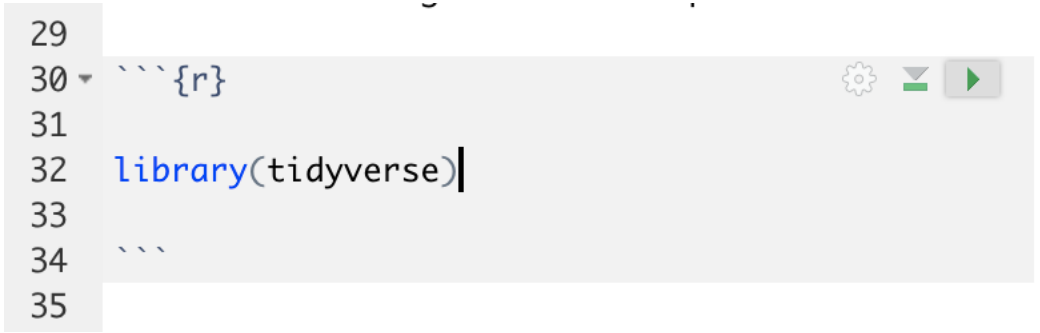
- Plots show up inline
- Easier to incorporate explanatory text and materials
- Like to be able to easily run one chunk at a time

Caution: Running things out of order can mean your code won't work again later. Clear your environment often and run code chunks in order to be safe.

R Markdown test

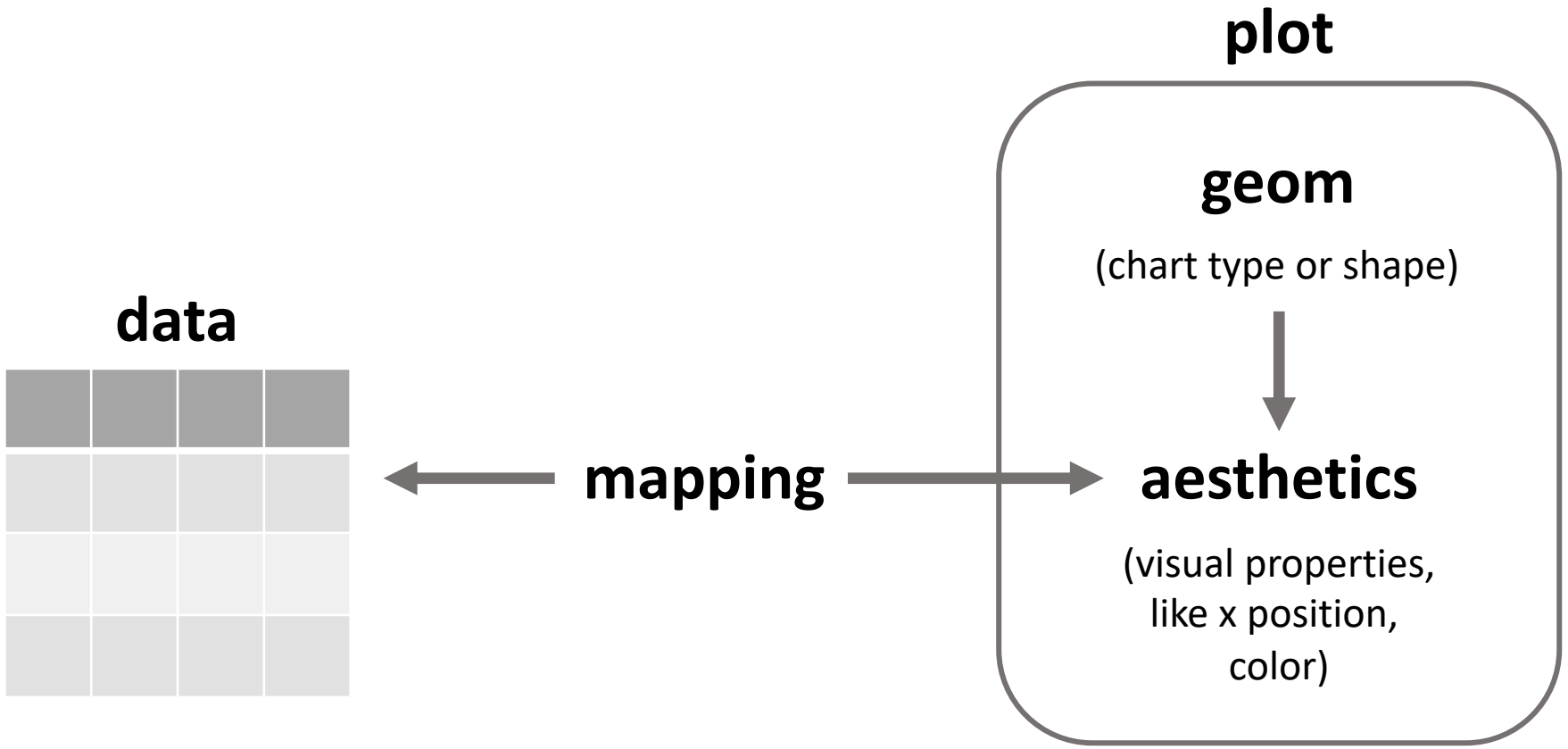
- File → New File → R Markdown
- Click OK to accept defaults
- Type inside the first few lines to edit the YAML header (edit title, add author, etc.)
- Add a new R code chunk at the end of the file using Insert → R
- Type some R code inside the code chunk:

  library(tidyverse)
- Run the new code chunk

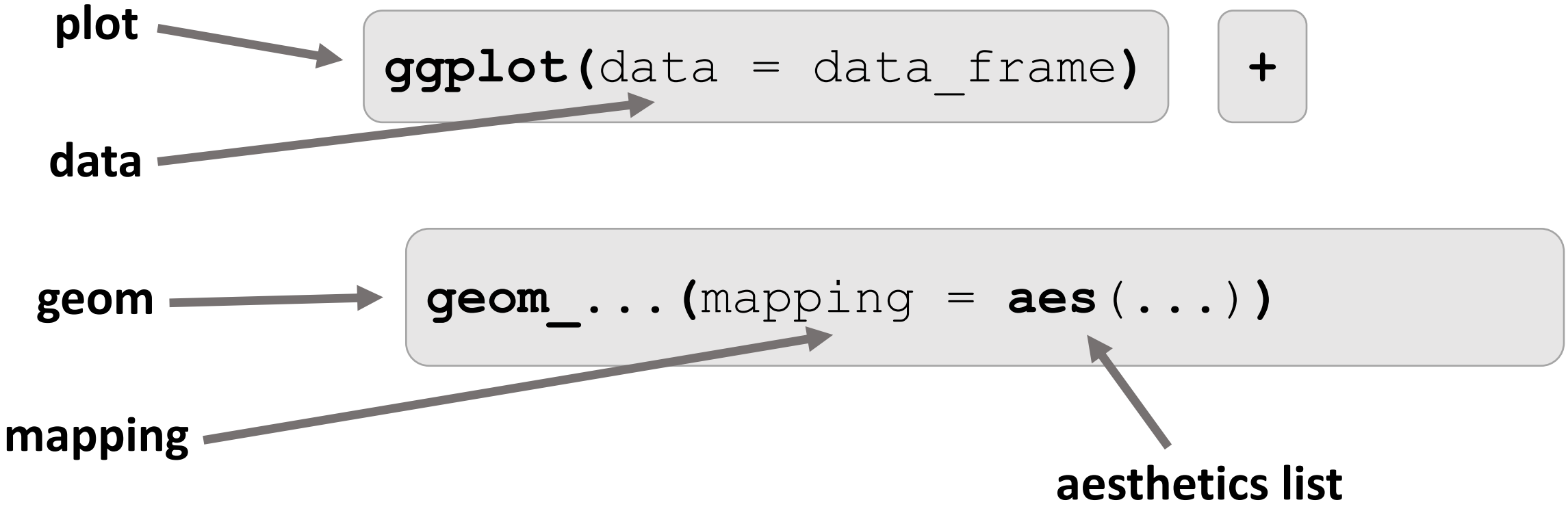


ggplot2: making a basic plot

Basic elements in any ggplot2 visualization



Template for a simple plot



1. Set the data

“iris”

Petal.Width	Petal.Length	Species
0.3	1.4	setosa
1.3	4.0	versicolor
2.1	5.7	virginica

```
ggplot(data=iris)
```

2. Choose a shape layer

"iris"

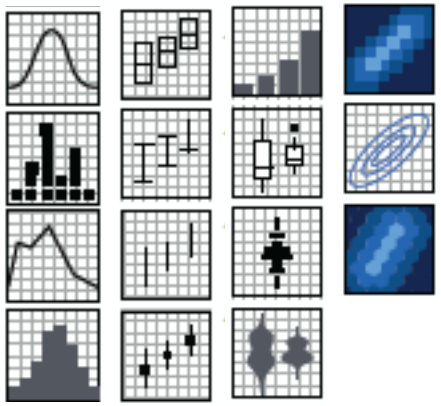
Petal.Width	Petal.Length	Species
0.3	1.4	setosa
1.3	4.0	versicolor
2.1	5.7	virginica

```
ggplot(data=iris) +
  geom_point()
```

```
Error: geom_point requires
the following missing
aesthetics: x and y
```

Types of geoms

- geom_bar()
- geom_point()
- geom_histogram()
- geom_map()
- etc.



ggplot2 cheatsheet

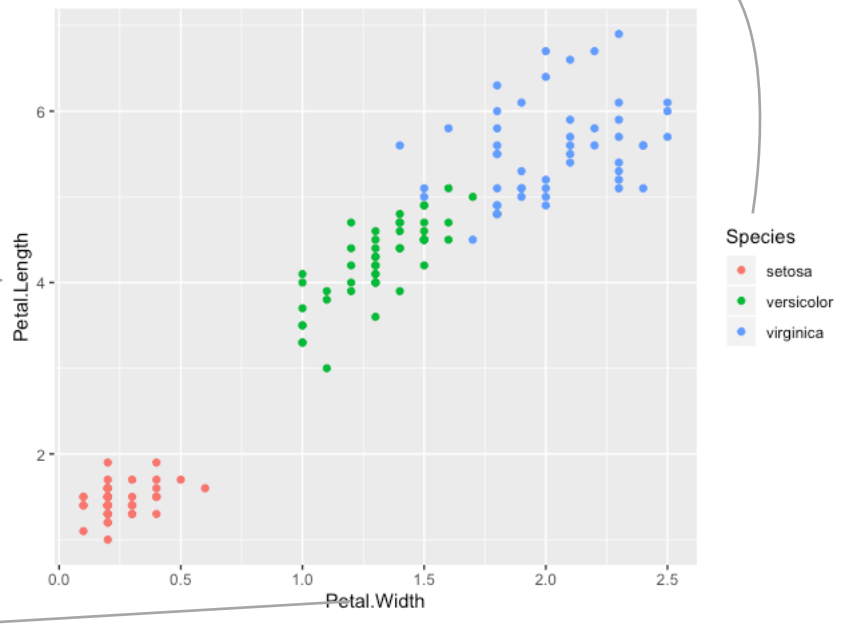
3. Map variables to aesthetics

"iris"

Petal.Width	Petal.Length	Species
0.3	1.4	setosa
1.3	4.0	versicolor
2.1	5.7	virginica

x position y position color

```
ggplot(data=iris) +
  geom_point(
    mapping=aes(x=Petal.Width,
                y=Petal.Length,
                color=Species))
```

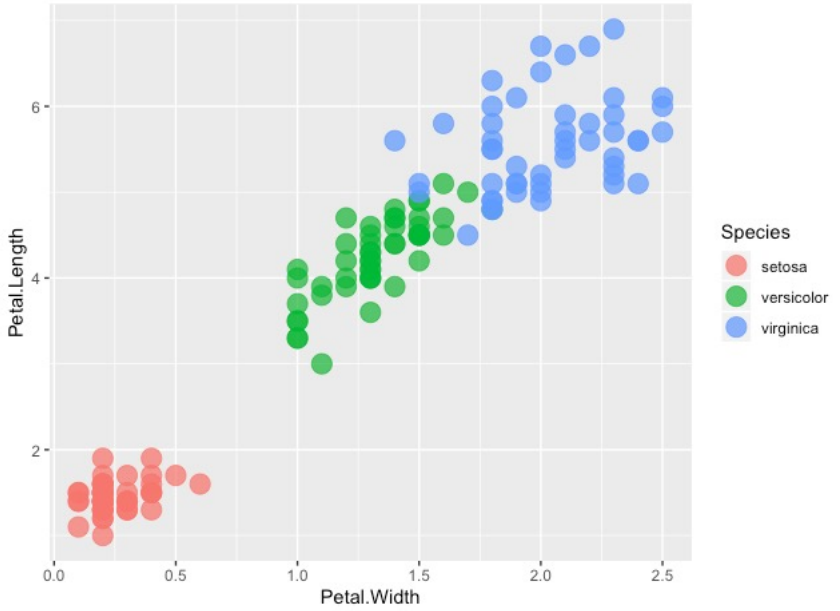


4. Add non-variable adjustments

"iris"

Petal.Width	Petal.Length	Species
0.3	1.4	setosa
1.3	4.0	versicolor
2.1	5.7	virginica

```
ggplot(data=iris) +
  geom_point(
    mapping=aes(x=Petal.Width,
                y=Petal.Length,
                color=Species),
    size=5, alpha=.75)
```



Fixing Errors

Debugging code

- Start simple
- If you see an error:
  - read error message for hints
  - check for problems with spelling/punctuation marks
- Get code to run without errors
- Check result to see if it makes sense

- Add a small change
- Get code to run without errors
- Check result to see if it makes sense
- etc.

RStudio built-in help documentation

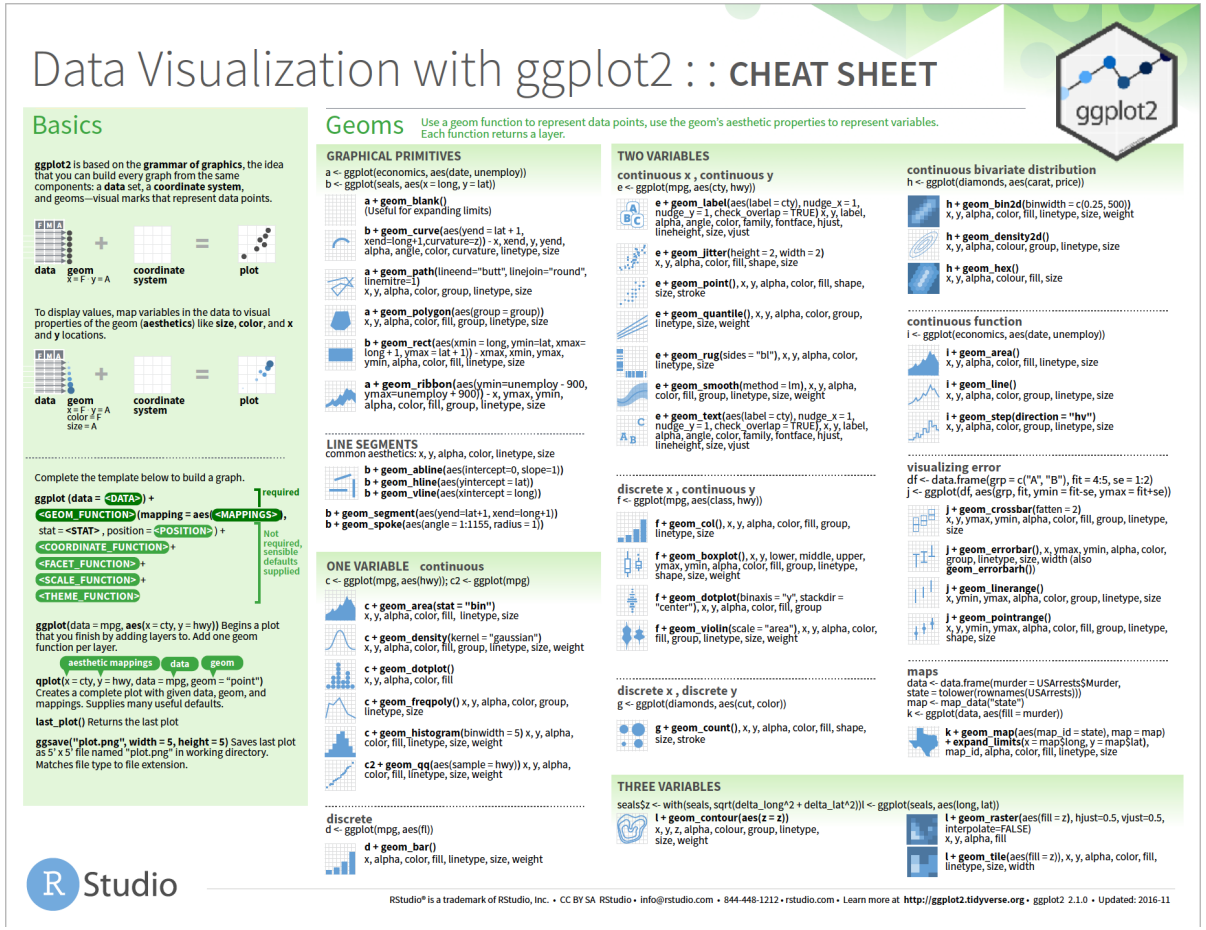
- In console, type
  `?<function or package name>`
- In help tab (not help menu), type into main search box
- In package tab, click on package name
- In help menu, use the Cheat Sheets submenu to download cheat sheet PDFs

ggplot2 Cheat Sheet

Help →

Cheatsheets →

Data Visualization with ggplot2



[ggplot2 cheatsheet]

Get workshop files

URL: https://github.com/amzoss/RVis-2Day

On GitHub:

- Click green “Code” button and select “Download ZIP”
- Unzip files on your computer
  - Windows: Double-click, then look for “Extract Files” at the top
  - Mac: Double-click
- Note: have noticed some issues when using OneDrive to store files

In RStudio:

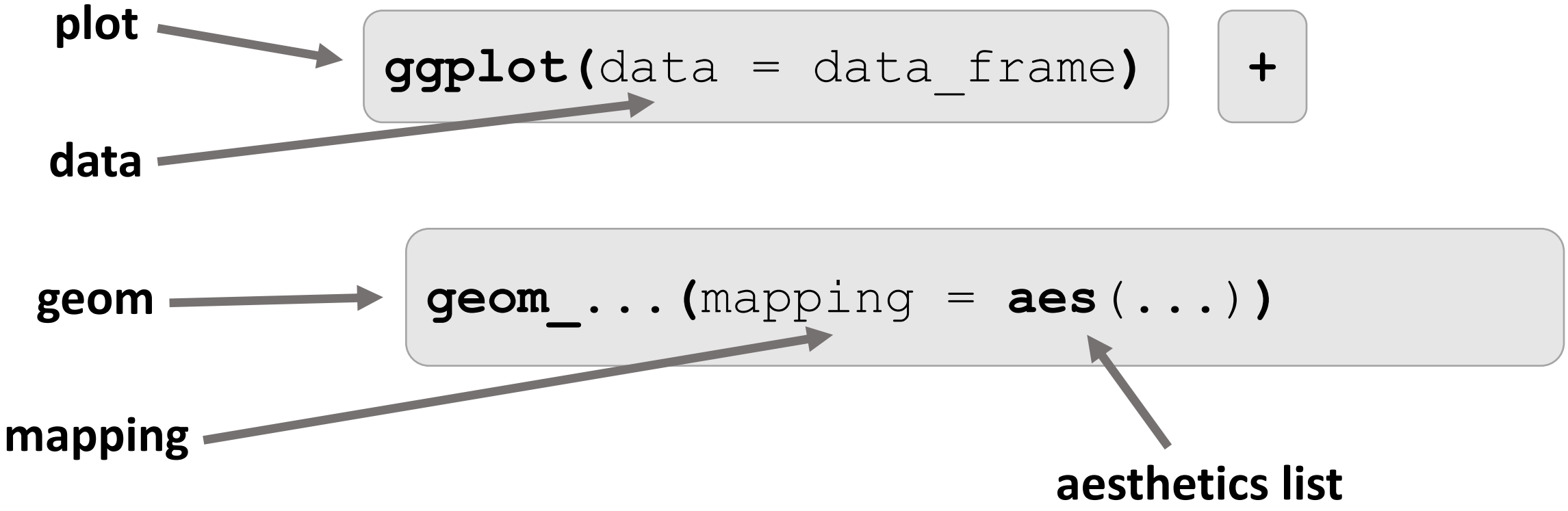
- Project → New project…
- Existing directory
- Select unzipped folder
- Create Project

Morning Break

Exercise 1: Inclusiveness Index

https://belonging.berkeley.edu/inclusiveness-index

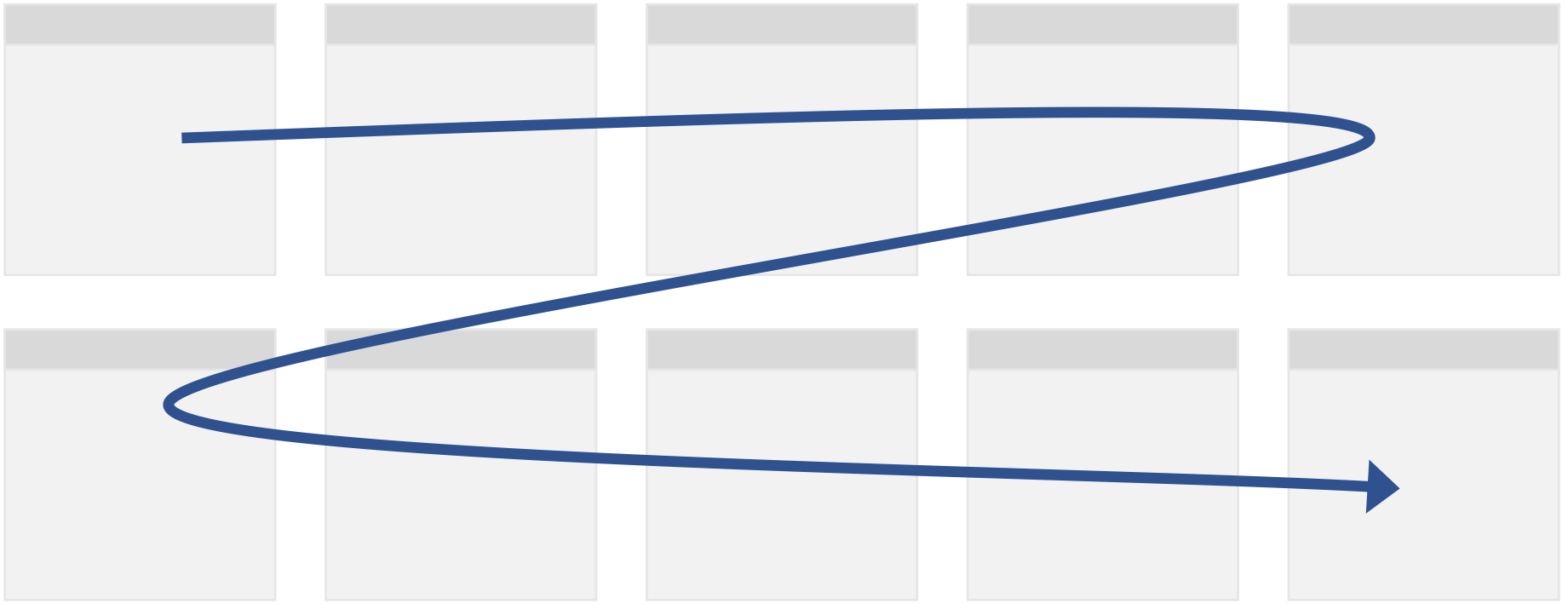
Template for a simple plot



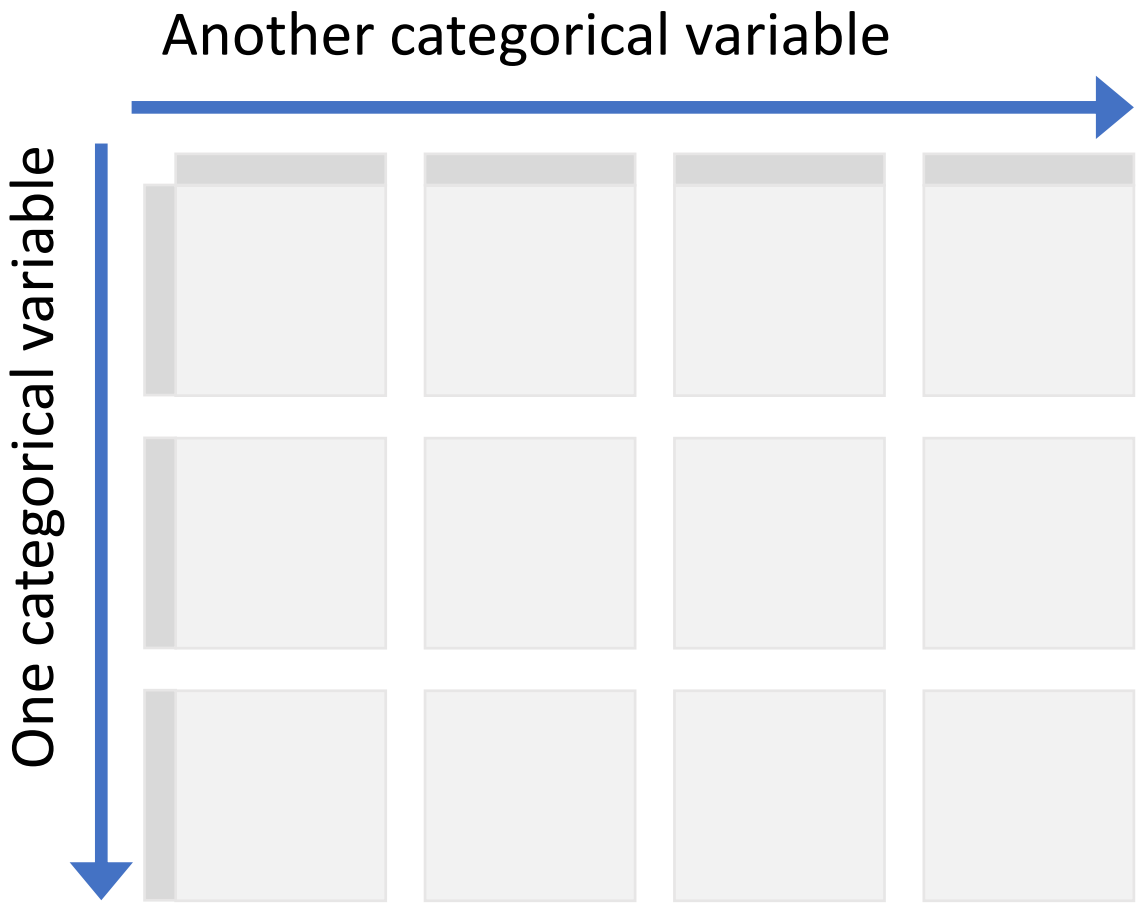
Creating repeated charts

facet_wrap()

```
+ facet_wrap(vars(variable))
```



facet_grid()

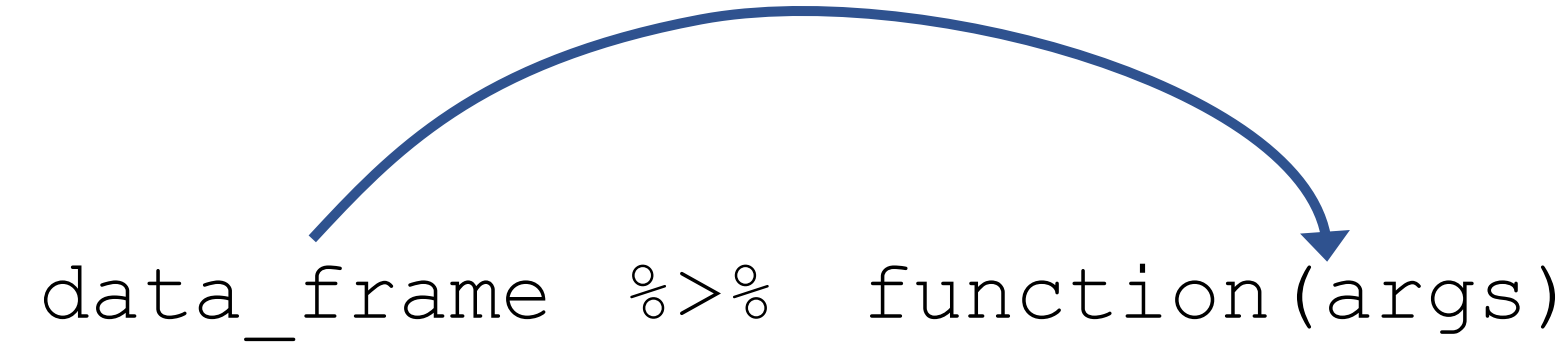



```
+ facet_grid(rows=vars(yvar),
               cols=vars(xvar))
```

Helpful data manipulation

Note: about %>%

- Loads automatically with tidyverse
- Used throughout tidyverse (except for ggplot2)
- Pushes data from the left into the function on the right



filter

Select a subset of rows

```
data %>% dplyr::filter(name == "John")
```

same as

```
dplyr::filter(data, name == "John")
```

dplyr cheatsheet

select

Select a subset of columns (many options!)

```
data %>% dplyr::select(id, name, age)
```

```
data %>% dplyr::select(-count)
```

dplyr cheatsheet

drop_na

Remove rows with NA values, either in any column or in specified columns

```
data %>% drop_na()
```

```
data %>% drop_na(age)
```

tidyr cheatsheet

count

Take a dataset, group it by one or more variables, and count the number of rows grouped. Count will be stored in a variable called “n”.

```
data %>% count(fruit)
```

fruit	n
apple	4
kiwi	10
orange	2

```
data %>% count(fruit, quality)
```

fruit	quality	n
apple	low	1
apple	high	3
kiwi	high	6
kiwi	medium	4
orange	Low	2

dplyr cheatsheet

count is same as group_by -> summarise

count() is shorthand for grouping by the categorical variable and then summarizing by the number of rows in each group.

```
data %>% count(fruit)
```

fruit	n
apple	4
kiwi	10
orange	2

```
data %>% group_by(fruit) %>%
  summarise(n = n())
```

fruit	n
apple	4
kiwi	10
orange	2

Pipe data into ggplot

When doing data manipulation, can be easier to pipe results to ggplot

```
data_frame %>% ggplot()
```

same as

```
ggplot(data = data_frame)
```

Lunch

Exercise 2: Customizing charts

Accessibility

All graphics need alternative text for screen reader users.

alt= "**Chart type** of **type of data** where **reason for including chart**"

Include a **link to data source** somewhere in the text

Note: Alt text should be relatively short. For longer descriptions, use add_description() from the savonliquide package

Writing alt text for data visualization/

Alternative Text in R and R Markdown

- ggplot2 now has alt option in labs(); gets read by shiny but not knitr
- in the meantime, use fig.alt in code chunk (just for HTML output)
  - can pull ggplot2 alt text into fig.alt with:
    ```` ```{r, fig.alt=ggplot2::get_alt_text(g)} g ``` ````
  - fig.cap will be used instead, if there is no fig.alt
- embedded images in the Markdown:

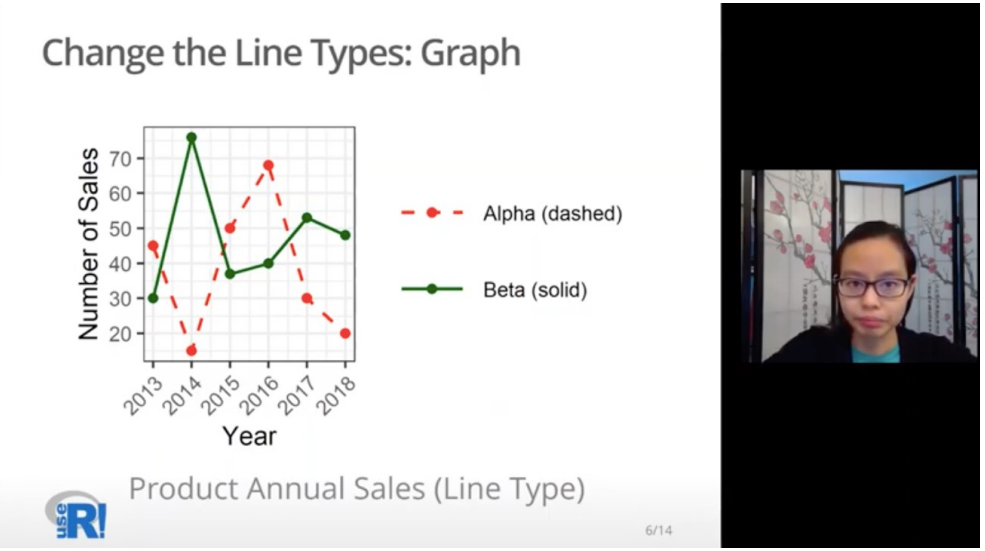
```
![text used for both alt text and figure caption](path/to/image)
```

Alt Text in R
```

Color Vision Deficiency

Use dual encoding (not just color)

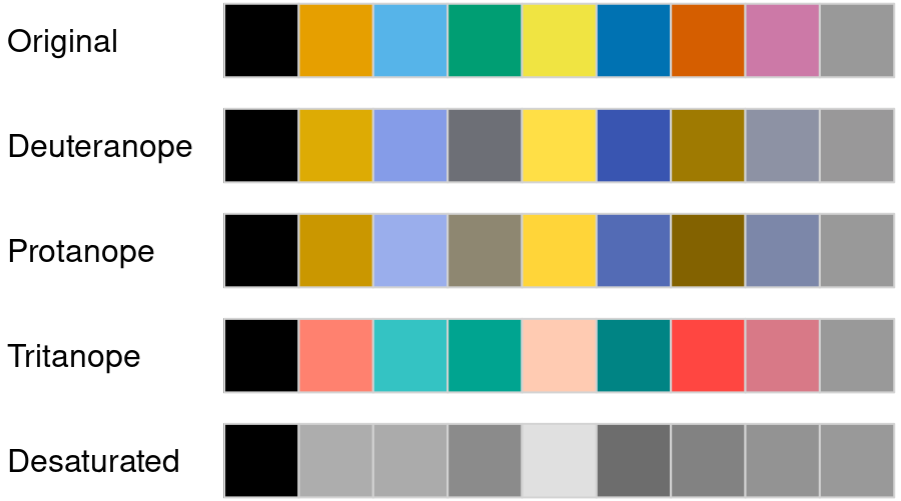
- Line color – also vary line type
- Point color – also vary point shape



[Improving accessibility in data visualizations created by ggplot2]

Use safe color palettes

- evaluate palettes to see how they look different for people with different types of color vision deficiency (CVD)



[colorspace package: CVD emulation]

Low Vision

- High color contrast
  - Both marks/text on background and labels on marks
  - Check contrast with savonliquide package
- Large text
  - See “output-examples.md” file for more sample code
  - Will cover in a later session

Converting graphics to sound, touch, text

- sonify package
- tactileR package
- BrailleR package
  - Note: set plot title, subtitle, caption using labs()

Accessible Data Science for the Blind Using R

Accessibility Resources

- savonliquide package
- Making betteR figures: Accessibility and Universal Design
- Highlights from the DVS accessibility fireside chat

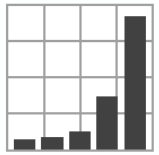
Scales

- Scales control how an aesthetics mapping displays in the chart, e.g.:
  - the labels that show up on the axis
  - the number of example sizes in a size legend
  - the colors used for a “fill” or “color” mapping
- Modify these properties by adding a scale layer to the chart

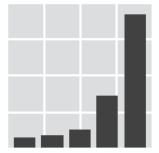
```
scale_x_continuous()
scale_y_log10()
scale_fill_discrete()
```

Themes

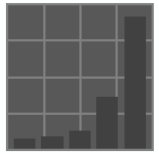
- Themes control properties of various visual elements, including:
  - Axis titles, text, ticks, lines
  - Plot colors, margins, text
  - Legend colors, margins, text
- Can add built-in themes as new layers, override specific theme elements, or build your own custom theme



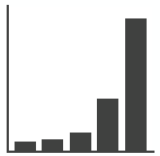
r + theme_bw()
White background with grid lines.



r + theme_gray()
Grey background (default theme).

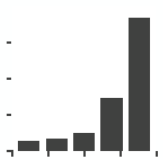


r + theme_dark()
Dark for contrast.



r + theme_classic()

r + theme_light()



r + theme_linedraw()

r + theme_minimal()
Minimal theme.

r + theme_void()
Empty theme.

ggplot2 cheatsheet

geom vs. scale vs. theme

Adding something that will appear
inside the **chart coordinate space**?

You will (almost always) be adding a **geom**!

Changing the way a **variable is displayed**?
(e.g., different axis breaks, different color mapping)

You will be adding a **scale**!

Changing the **look and feel** of the chart?

You will be adding or making changes to a **theme!**

More practice: Advanced ggplot2 workshop

Workshop video

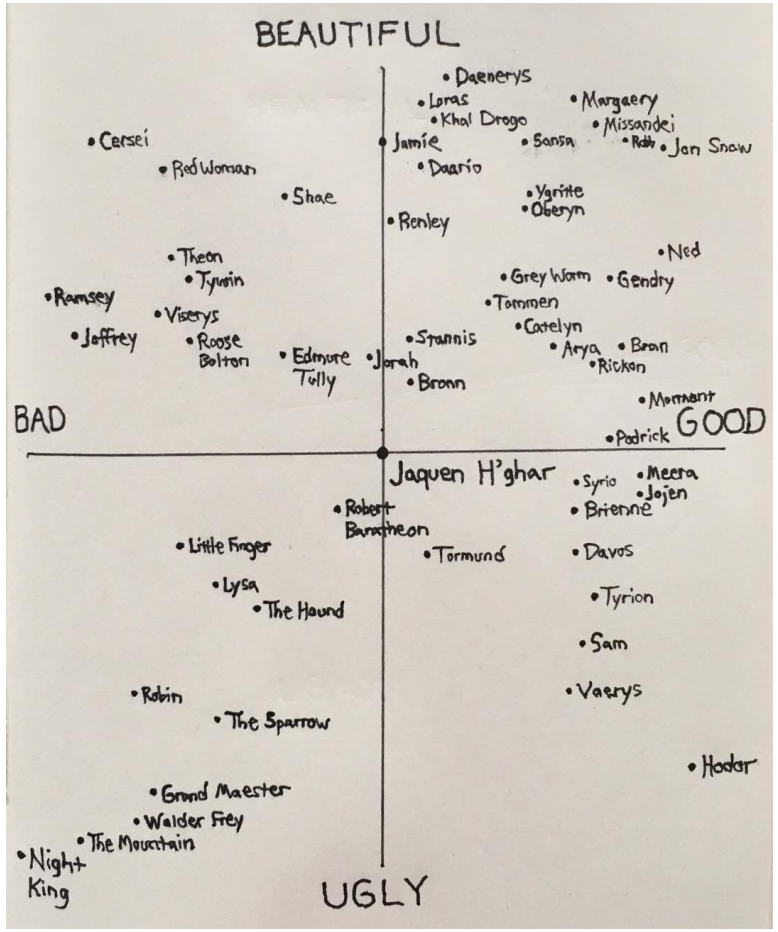
Workshop materials

Afternoon Break

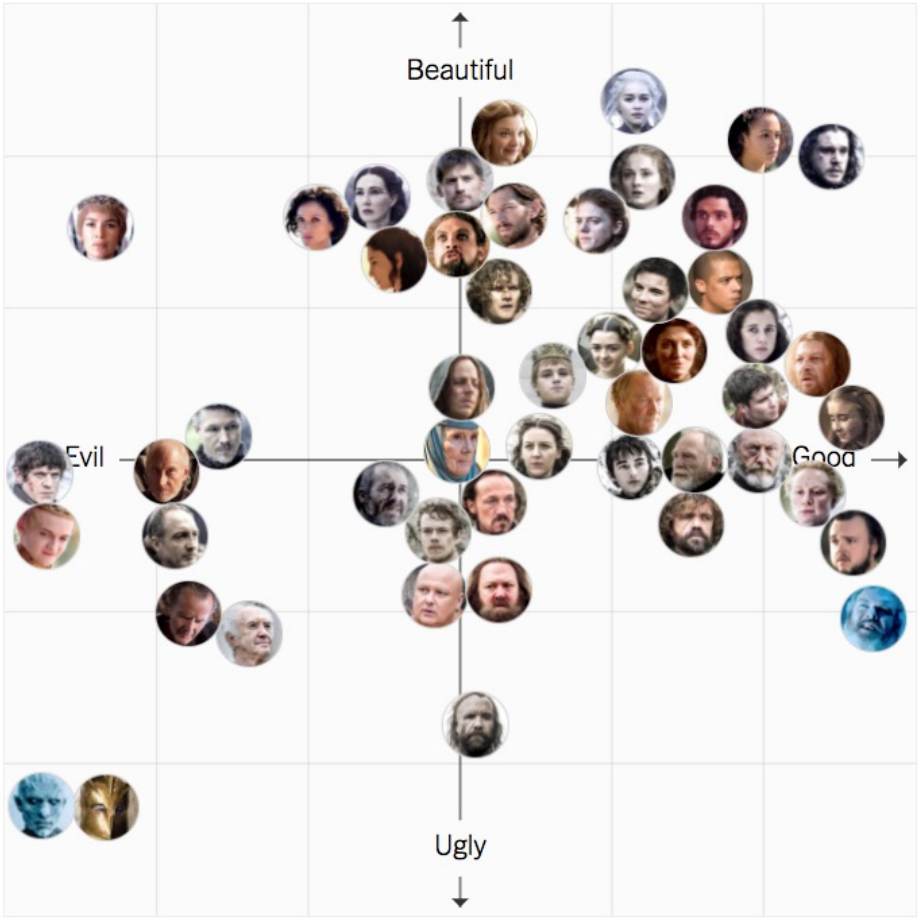
Exercise 3:
Game of Thrones character ratings

https://www.nytimes.com/interactive/2017/08/09/upshot/game-of-thrones-chart.html

Game of Thrones character ratings



https://www.instagram.com/p/BWnn-YogX1n/



https://www.nytimes.com/interactive/2017/08/09/upshot/game-of-thrones-chart.html

ggplot2: inheritance

Template for a simple plot (review)

main plot function

```
ggplot(data = data_frame)  +
```

shape layer

```
geom_...(mapping = aes(...),
              non-variable adjustments)
```

Expanded template

main plot function

```
ggplot(data = data_frame,
       mapping = aes(...))
```

+

shape layer

```
geom_...(data = data_frame,
         mapping = aes(...),
         non-variable adjustments)
```

Inheritance

data and aesthetics will carry through from main function to shape layers

main plot function

```
ggplot(data = data_frame,
       mapping = aes(...))
```

+

shape layer

```
geom_...(data = data_frame,
         mapping = aes(...),
         non-variable adjustments)
```

+

shape layer

```
geom_...(data = data_frame,
         mapping = aes(...),
         non-variable adjustments)
```

Other helper packages

- gganonymize to randomize text in ggplot2 figures
- visdat to visualize variable classes and missing data
- ggthemes for additional themes and scales, especially ones that match software defaults (e.g., Tableau)
- esquisse for building ggplot2 charts interactively
- colorblindr for simulating color vision deficiency
- ggpubr for publication-ready plots

ggplot2 Resources

- General ggplot2 information
- R Graphics Cookbook (recipes for plots)
- R for Data Science (online book that includes ggplot2)
- ggplot2: Elegant Graphics for Data Analysis (book by Hadley Wickham)
- ggplot2 cheatsheet (also in RStudio)
- Data Carpentry lesson on ggplot2
- Data Visualization: A Practical Introduction, by Kieran Healy
- RStudio “Visualize Data” Primer

Thanks for your feedback!

angela.zoss@duke.edu