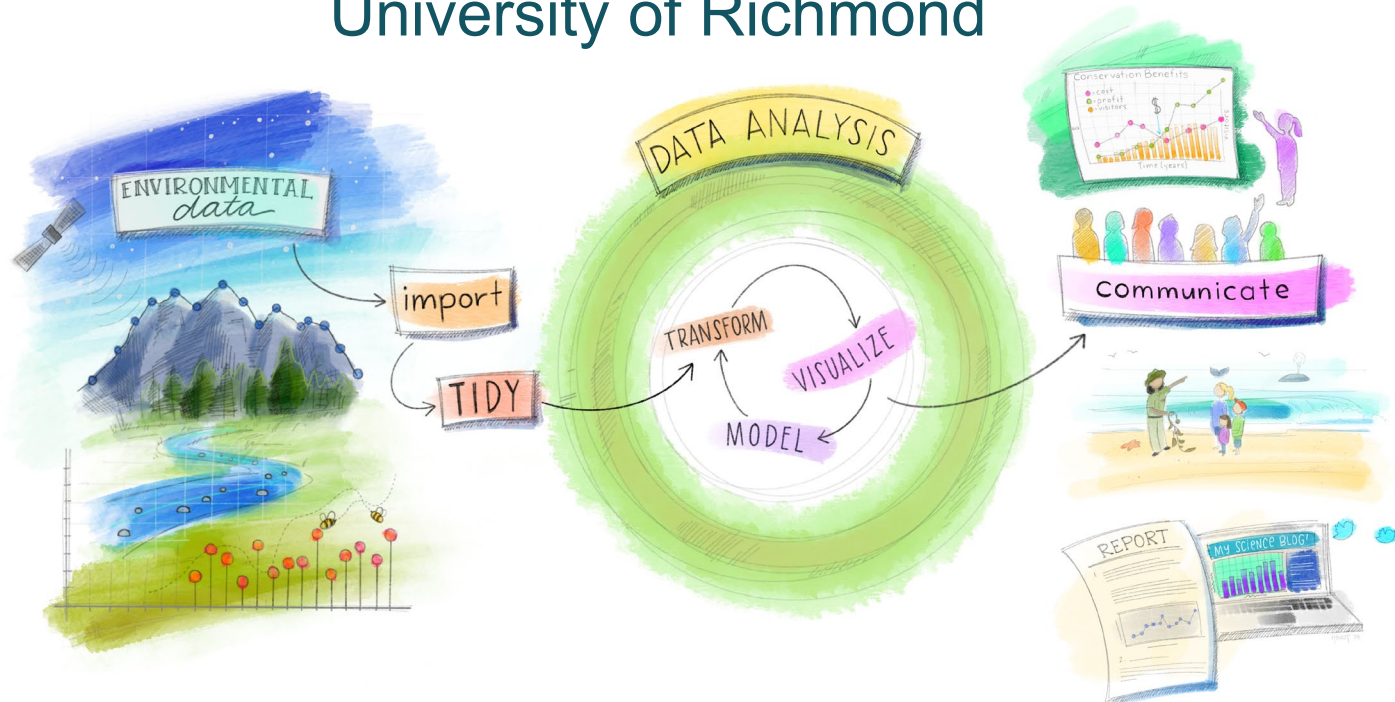


Teaching biology students exploration and visualization in a data-driven world

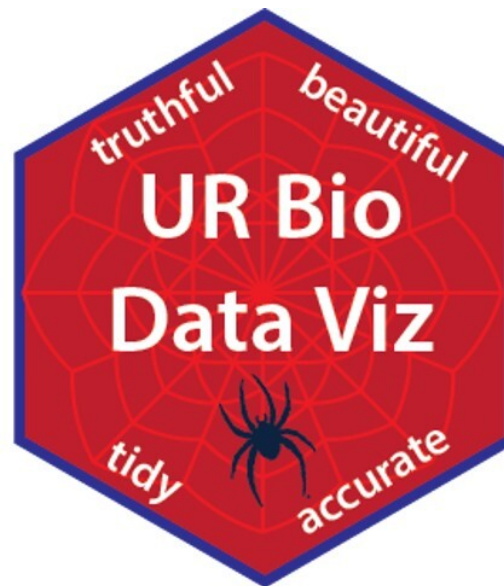
Dr. Kristine Grayson & Dr. Angie Hilliker
University of Richmond

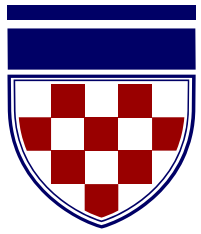


Updated from Grolemund & Wickham's classis R4DS schematic, envisioned by Dr. Julia Lowndes for her 2019 useR! keynote talk and illustrated by [Allison Horst](#).

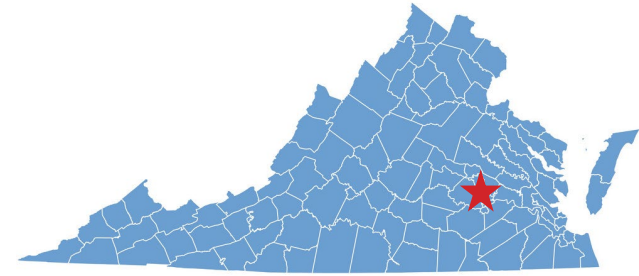
Many of our course materials are available on QUBES

- Syllabus and Resources (including these slides):
- <https://qubeshub.org/publications/2450/1>
- Assignments and Projects:
- <https://qubeshub.org/publications/2452/1>





UNIVERSITY OF RICHMOND



4,002

Total enrollment (3,147 undergrad, 855 grad students)

16

Average number of students per
undergraduate class

22nd

—U.S. News & World Report

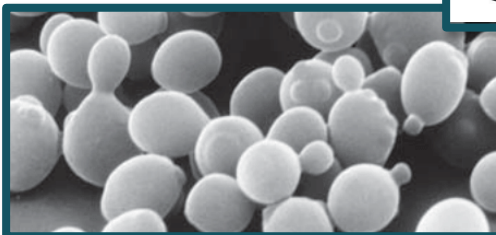
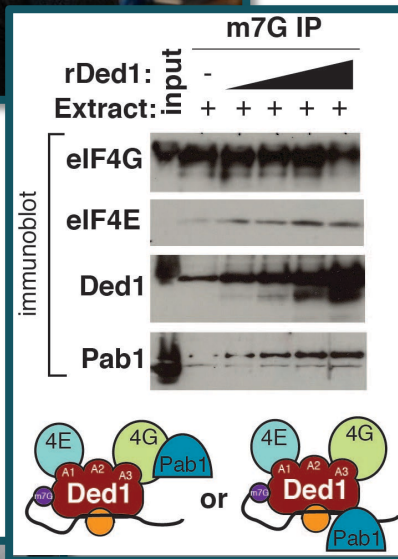


30%

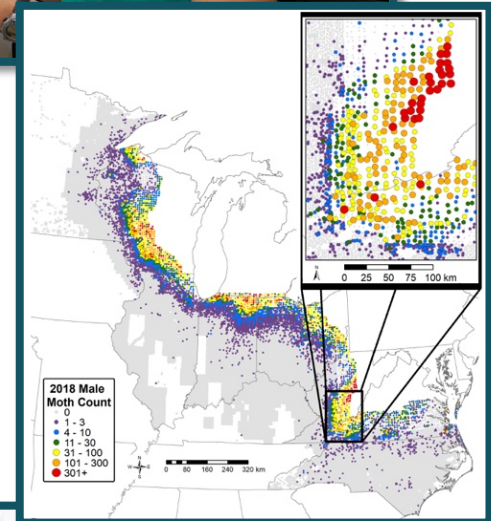
Arts & Sciences
students conduct
research



Angie Hilliker



Kristine Grayson



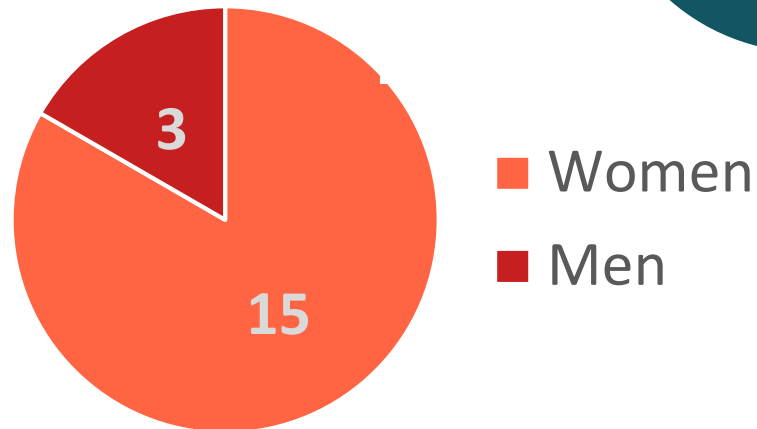
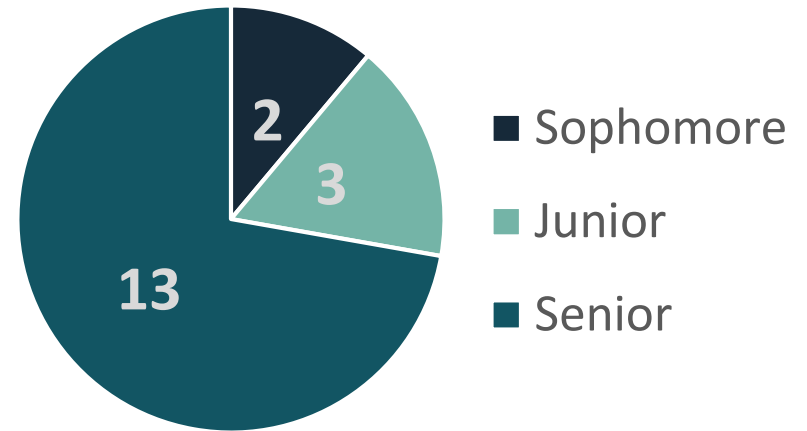
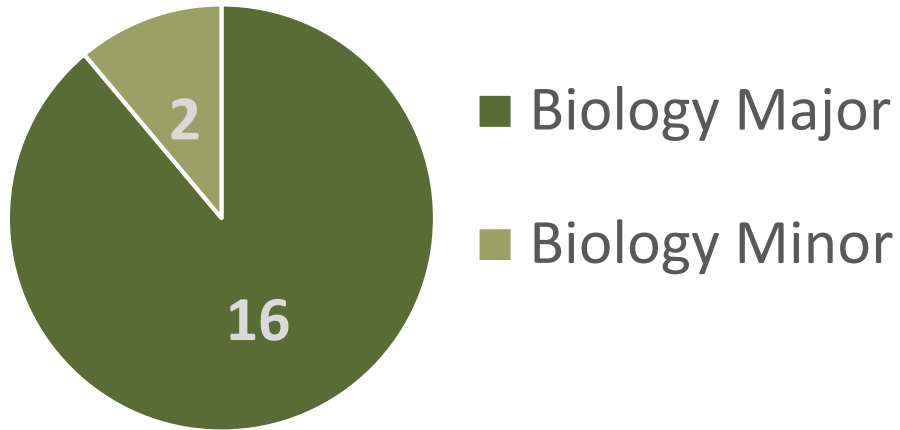
Student experience before our class

Biology Curriculum:

- Intro to Biological Thinking
- Integrated Biological Principles I and II
- Upper level biology electives

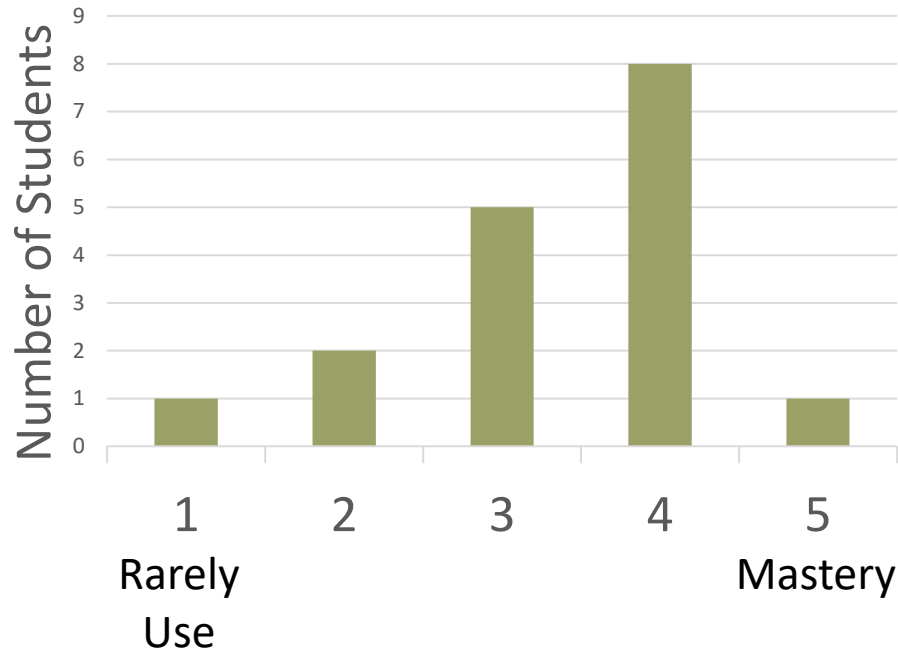


Class Demographics

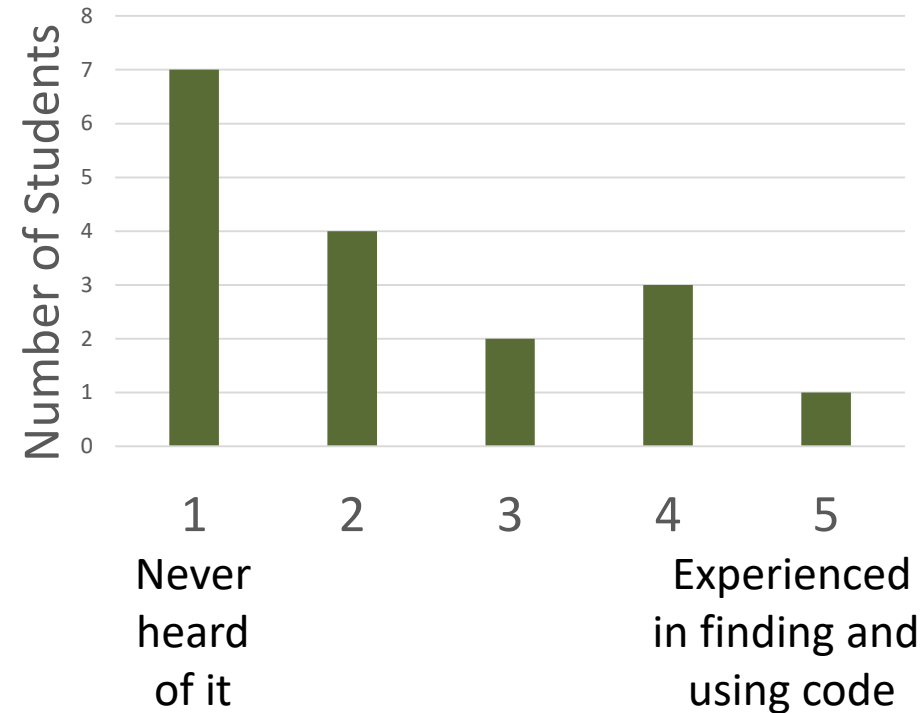


Students' Prior Programing Experience

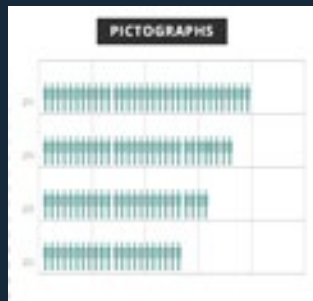
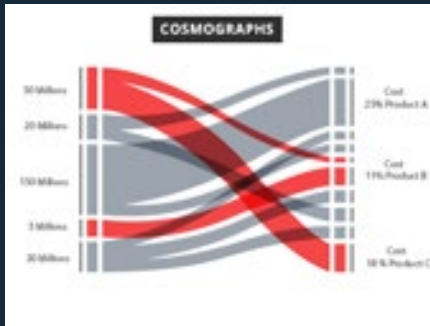
Comfort Level with Excel



Comfort Level with R



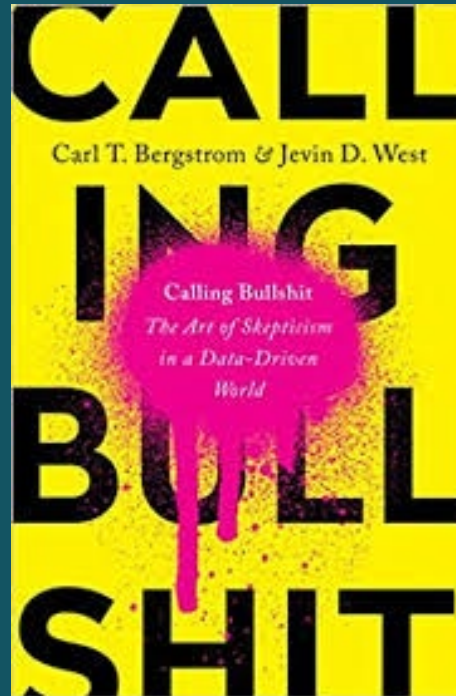
Data Visualization Principles



Accurate?
Clear?
Beautiful?

typesofgraphs.com

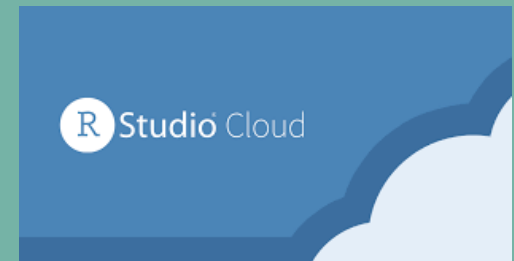
Critical Thinking or “Calling Bullshit”



How to see
manipulations,
flaws, and fraud.

by Carl T. Bergstrom & Jevin D. West

Tools and Tricks



Our chosen tools

Students will be able to:

Visualize

Learn to select appropriate visualizations for a variety of data types

Identify and avoid common pitfalls that can result in inaccurate presentation of data and figures

Convey clear, accurate, beautiful graphics with data

Manage

Acquire and evaluate publicly available data

Clean, organize, and manipulate data sets

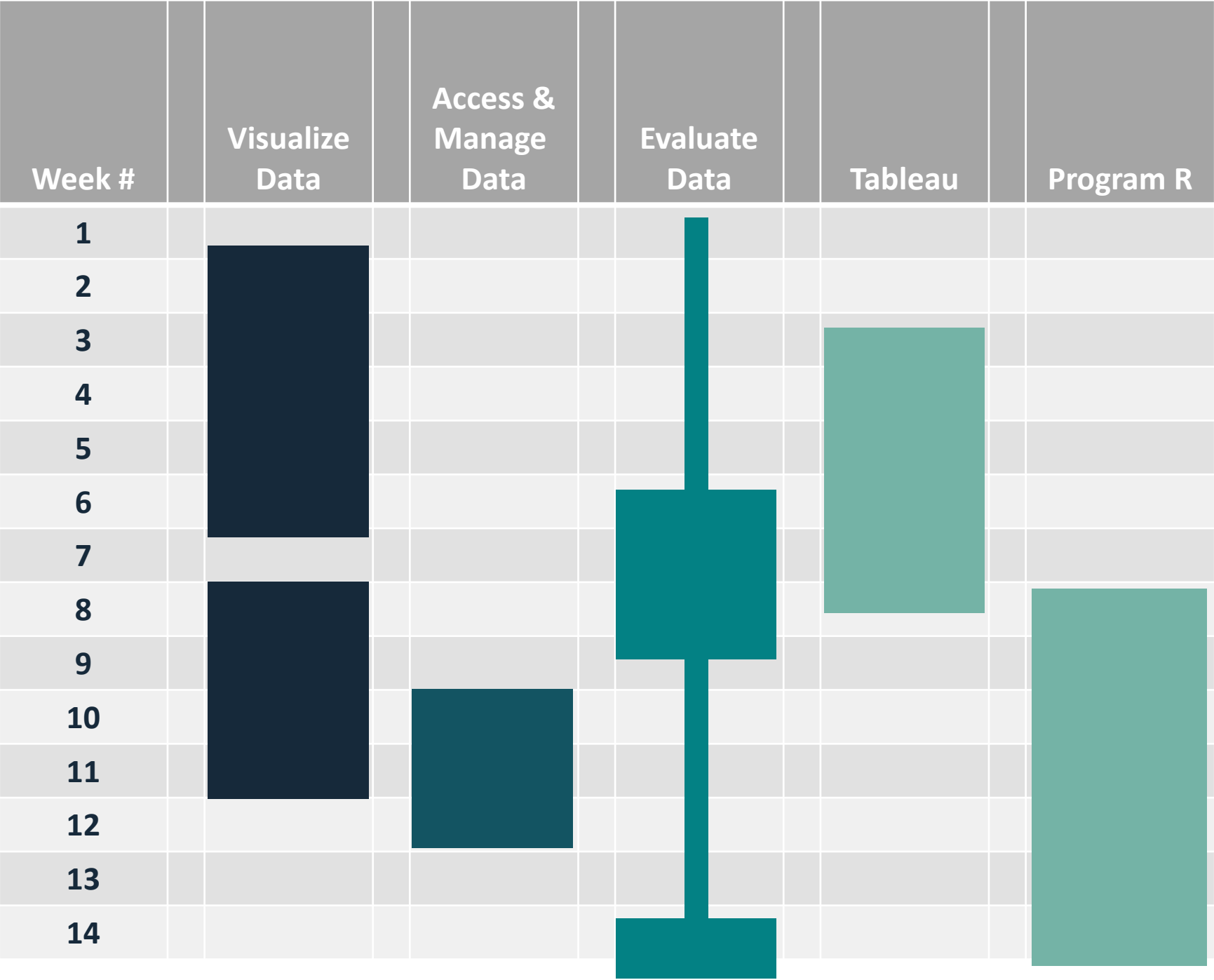
Evaluate

Recognize misuse of data, such as “p-hacking” and data dredging

Learn to recognize misinformation and BS

Tools

Use Tableau & Program R as software tools for creating visualizations

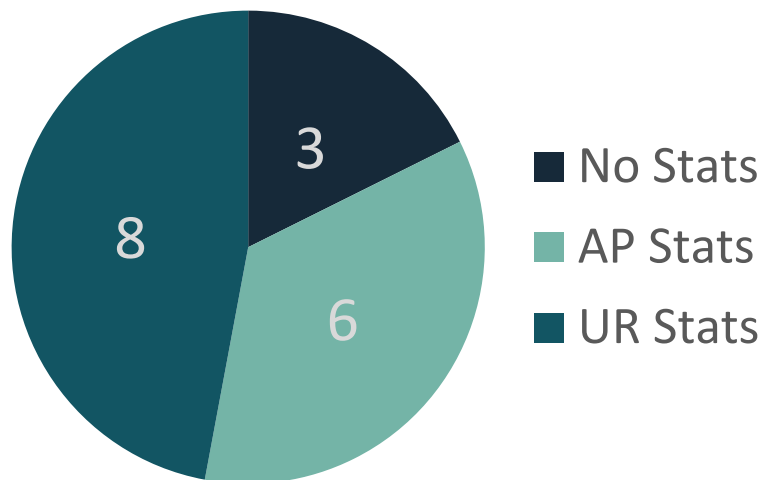


Data Communication topics :



- Best practices for exploratory data analysis and contrast with hypothesis testing
- Scientific process and P value hacking
- Reproducible research and open data
- Data ethics and inclusion

What we aren't teaching in this class:



What we aren't teaching in this class:



```
1 say_hello <- function(name) {  
2   message <- paste("Hello", name)  
3   return(message)  
4 }
```


Tableau and Program R, basics



- GUI interface
 - Tableau Online
 - Good job skill in the business world
 - Instructors and students can get educational license
 - Good online tutorials and guides
- R Studio provides a partial GUI interface
 - R Studio Cloud
 - Used by many scientists
 - Open source, but R Studio Cloud has limited access without paid plan
 - Excellent (and prolific) online tutorials and guides

Rational for Tableau and Program R



- Explore Data Viz principles without code
 - Grouping/filtering, data cleaning, and calcs are easy to learn
 - Quickly change between graph types/styles
 - Handles large data sets
 - Dashboards
- One way to introduce students to programming
 - Powerful packages for data analysis & viz
 - Exceptionally wide range of graphic types available
 - Can handle exceptionally large data sets
 - Knitting and Markdown

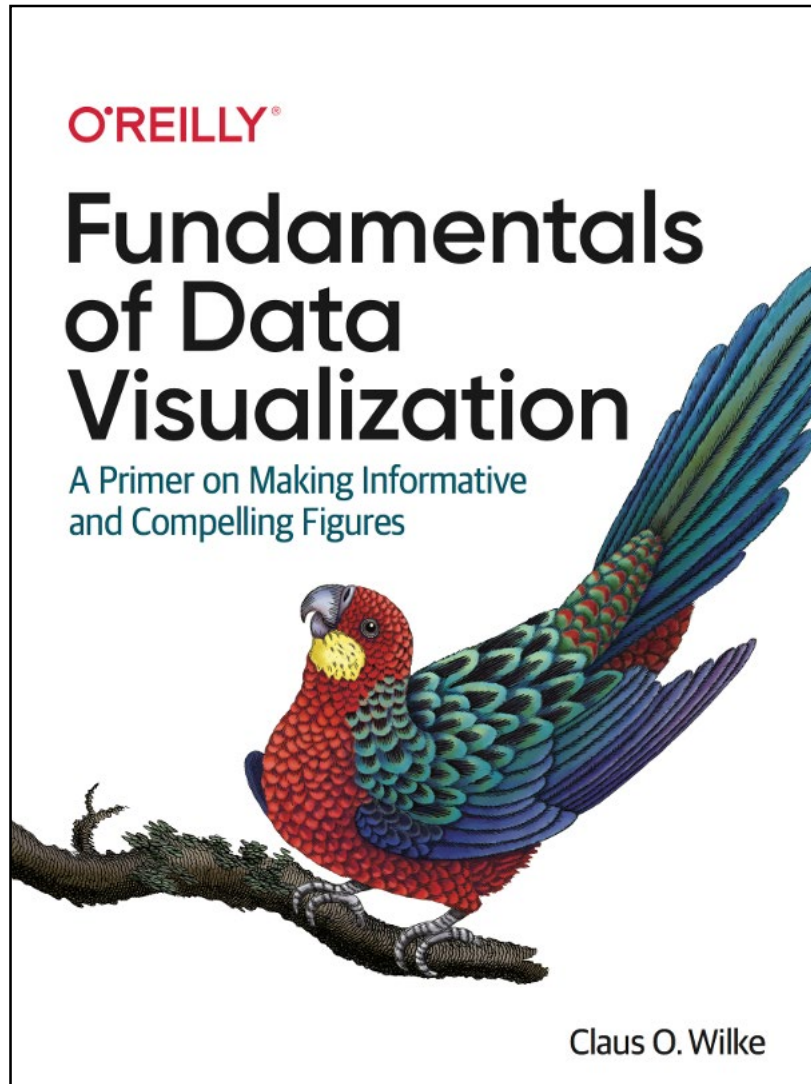
Drawbacks for Tableau and Program R



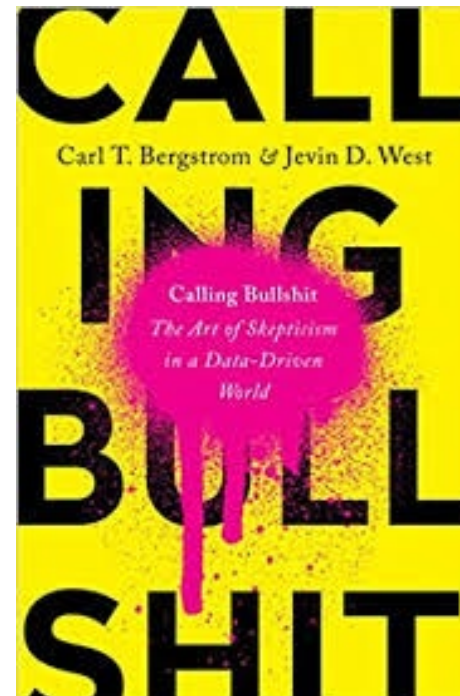
- Limited by the catalogue of graph types available
- Limited analysis/stats
- Error bars are difficult
- Lots of new terminology specific to Tableau
- Have to introduce programming principles
- Steep learning curve
- Packages have documentation of varying quality

Formative Assignments/Activities

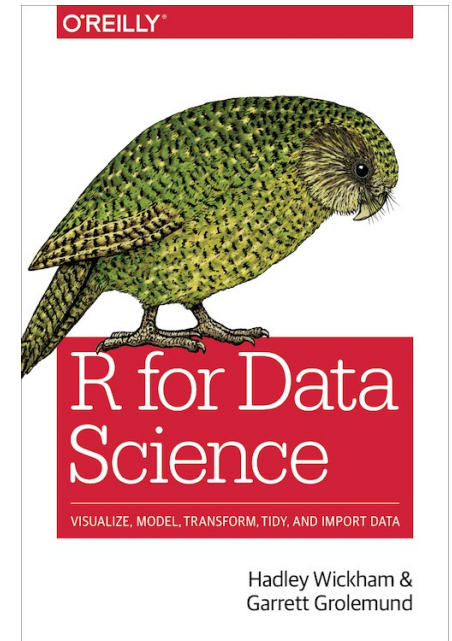
Class Reading



[Open Source](#) Copy



www.callingbullshit.org



[Open Source](#) Copy

Perusall promoted conversations between students & with instructors



graded error bars. The grading helps the reader perceive that there is a range of different possibilities. If I showed simple error bars (without grading) to a group of people, chances are at least some of them **would perceive the error bars deterministically**, for example as representing minimum and maximum of the data. Alternatively, they might think the error bars delineate the range of possible parameter estimates, i.e., the estimate could never fall outside the error bars. These types of misperception are called *deterministic construal errors*. The more we can minimize the risk of deterministic construal error, the better our visualization of uncertainty.

Error bars are convenient because they allow us to show many estimates with their uncertainties all at once. Therefore, they are commonly used in scientific publications, where the primary goal is usually to convey a large amount of information to an expert audience. As an example of this type of application, Figure 16.7 shows mean chocolate ratings and associated confidence intervals for chocolate bars manufactured in six different countries.



Reading Assignment

✓ Upvoted by instructor +1 ✓

LP This is a great point-I had never thought about how error bars could be perceived this way before. I have also never seen the graded error bars before. I am not sure if I have ever seen error bars in any context besides a scientific journal article. However, I do think they convey important information, so maybe the graded error bars could be included in figures meant for the general public such as magazine figures.

Feb 11 3:24 pm

KG I was also thinking about how infrequent error bars are shown outside of journal publications and why had I never noticed that before?!

Feb 14 10:36 pm

AA I was thinking the same thing, Lindsay. As someone regularly exposed to visualizations in the science field, I forget how the general public may perceive certain things much differently than I do. I like the idea of using graded error bars to make the perception of the figure clearer.

Feb 15 11:22 am

Comments &
Discussion

Perusall allowed students to ask questions:

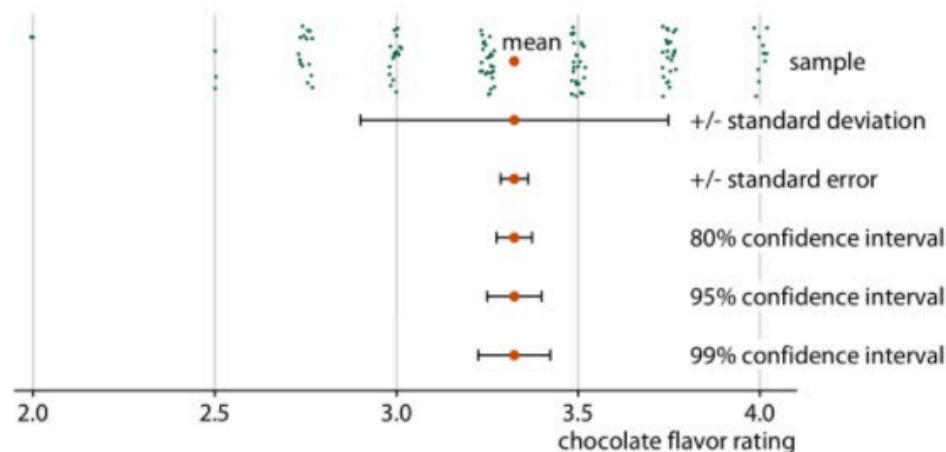
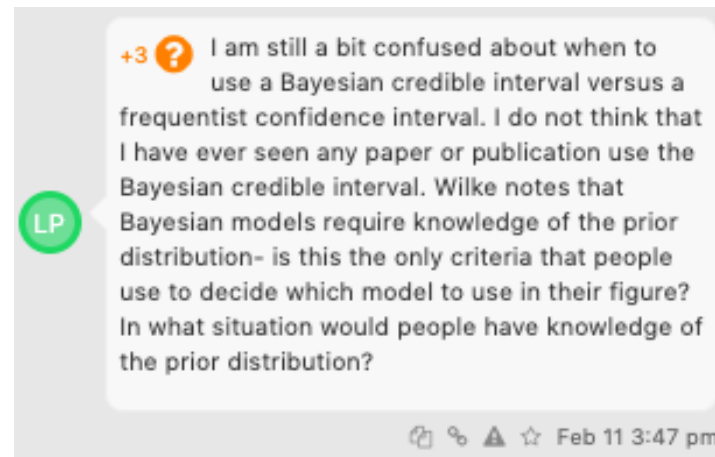
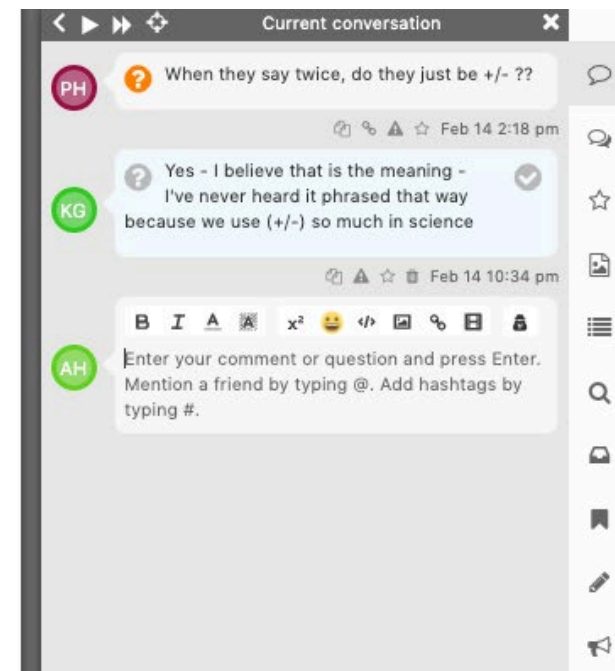


Figure 16.5: Relationship between sample, sample mean, standard deviation, standard error, and confidence intervals, in an example of chocolate bar ratings. The observations (shown as jittered green dots) that make up the sample represent expert ratings of 125 chocolate bars from manufacturers in Canada, rated on a scale from 1 (unpleasant) to 5 (elite). The large orange dot represents the mean of the ratings. Error bars indicate, from top to bottom, twice the standard deviation, twice the standard error (standard deviation of the mean), and 80%, 95%, and 99% confidence intervals of the mean. Data source: Brady Brelinski, Manhattan Chocolate Society



Perusall showed us what resonated with students



MH This is the first time I have even partially understood log scales and why to use them, and this sentence puts it more clearly than anything else I've seen! I hope that throughout this course we interact a bit more with log scales and I will be able to wrap my brain around them :) **+3**

ZY The explanation here for the difference is very clear. I did make related mistakes before when I made figures. It's an important point that is easy to neglect. **+2** ✓

📄 🔗 ⚠️ ☆ Feb 12 9:44 am

MH ? I agree, I think this is the most succinct explanation I've ever read! I also frequently forget the difference between standard error and standard deviation, but it seems like standard deviation is best for comparing across groups. ✓

📄 ⚠️ ☆ Feb 13 5:48 pm

Perusall: encouraged students to look beyond the reading and share what they discovered!



I'm also curious if anyone has seen a graph with the specific axes mentioned in the text though!

LP

I thought it was strange that there was no data for 2012, so I decided to do some research. I found that David Pérez García became mayor of Alcorcón at the end of 2011. There were conflicting sources, but most seemed to indicate that the unemployment rate rose quite a bit during 2012. It is interesting that, in addition to the scaling flaws discussed in this chapter, this data appears to be cherry-picked as well to make it seem that the current mayor had a larger impact than he truly did.

+4



LP

I googled CVD simulators and found that some of the colors that I had used in creating my figures in the past were unable to be read by people with CVD. I am sad to say that I had never thought about this when creating my figures before. I am really glad to know that there are color scales such as the one used by this book that allow figures to be understood by nearly everyone. I also found websites where people with CVD can upload images and they would convert the colors into ones that they can see. However, I still think it is important to not put this burden on people and instead we should create figures that are accessible to everyone.

+4

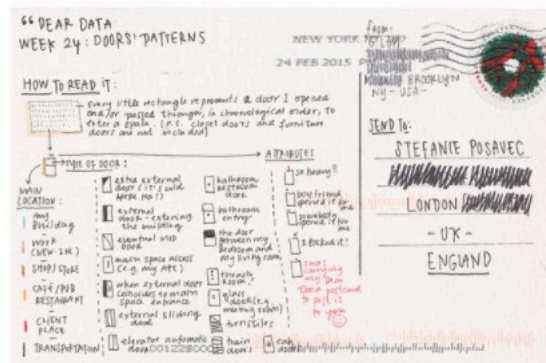




Week Twenty-four



Stefanie

[illegible]

dear-data.com

Dear Data Assignment



Prompt Option 2:
#08 WHAT MY CAMERA SEES

What do you normally train your camera on?

Open the photo album on your phone and/or computer.

Choose a time period you want to survey (there's room for 360 photos), and add your dates to the drawing.

ONE WEEK?
ONE MONTH?
THREE MONTHS?

Each photo is represented by a COLOR and a SYMBOL.

1. COLOR = the location of the photo

2. The SYMBOL = the type of photo you took

3. Connect multiple photos taken of the same thing

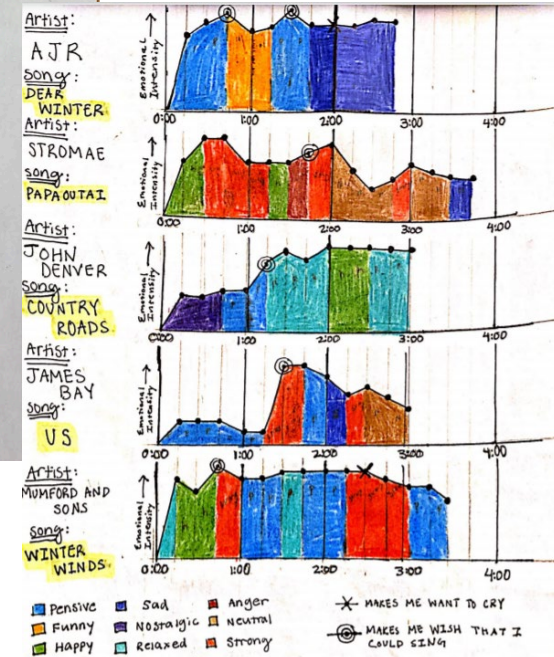
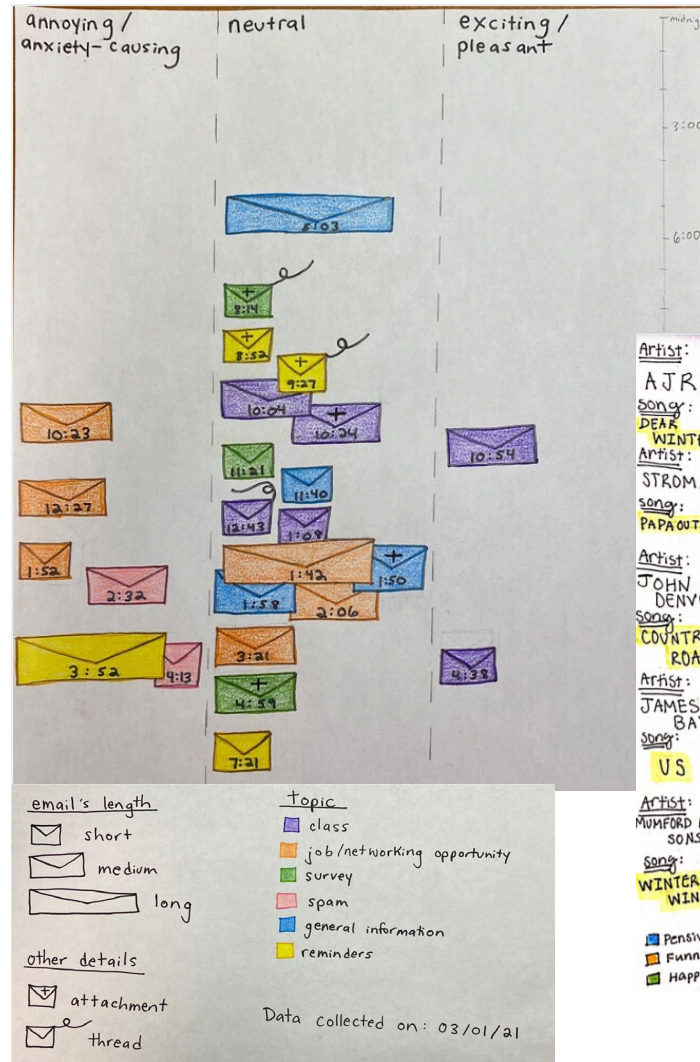
4. Underline the photos actually worth keeping for the future!

5. Add additional notes in the margins near the photos.

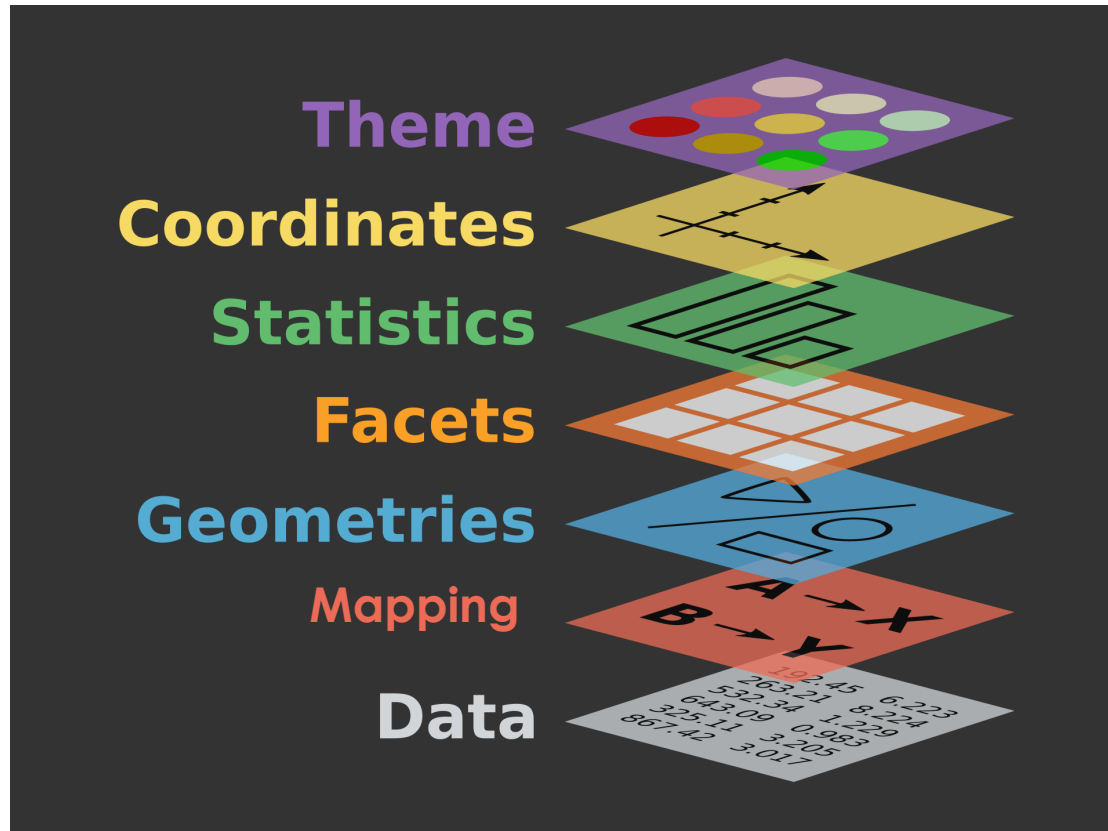
Photos were taken from _____ to _____

9

- Representative work of my favorite painter
- What does my clothes collection say about me?
- How music makes me feel



Grammar of Graphics: Framework of ggplot2 in R



[*The Grammar of Graphics*](#) by Leland Wilkinson (2005)

[*A Layered Grammar of Graphics*](#) by Hadley Wickham

(2010)

What we taught in R

Tidyverse

Packages Blog Learn Help Contribute



R packages for data science

The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

<https://www.tidyverse.org/>

Live Demos in Class with R Markdown (.rmd) files



```
---  
output: html_document  
---
```

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
```{r}  
summary(cars)
```
```

You can also embed plots, for example:

```
```{r, echo=FALSE}  
plot(cars)
```
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

R Markdown

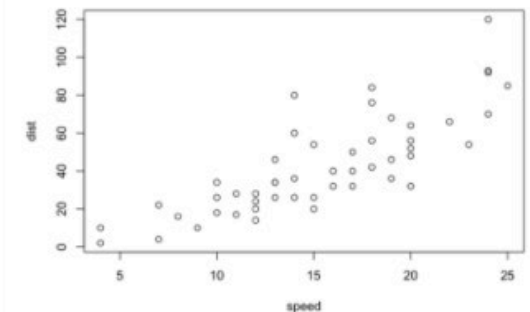
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

| ## | speed | dist |
|----|--------------|-------------|
| ## | Min. : 4.0 | Min. : 2 |
| ## | 1st Qu.:12.0 | 1st Qu.: 26 |
| ## | Median :15.0 | Median : 36 |
| ## | Mean :15.4 | Mean : 43 |
| ## | 3rd Qu.:19.0 | 3rd Qu.: 56 |
| ## | Max. :25.0 | Max. :120 |

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Live Demos in Class with R Markdown files



The pipe (`%>`, read: "and then")

When writing complex data analysis pipelines, we frequently use the pipe operator `%>` to move data from one analysis step to the next. The pipe is pronounced "and then", and it takes the data on its left and uses it as the first argument for the function on its right.

For example, to see the first few lines of a dataset, we often write `head(data)`. Instead, we can write `data %>% head()`.

Try this yourself. Write code that displays the first few lines of the `penguins` dataset, using `%>` and `head()`.

Any guesses on how one can see the last few rows of the data frame?

```
```{r pipe-solution}
```

```
| ...
```

Instructor provides description and creates an empty “code chunk.”

## The pipe (`%>`, read: "and then")

When writing complex data analysis pipelines, we frequently use the pipe operator `%>` to move data from one analysis step to the next. The pipe is pronounced "and then", and it takes the data on its left and uses it as the first argument for the function on its right.

For example, to see the first few lines of a dataset, we often write `head(data)`. Instead, we can write `data %>% head()`.

Try this yourself. Write code that displays the first few lines of the `penguins` dataset, using `%>` and `head()`.

Any guesses on how one can see the last few rows of the data frame?

```
```{r pipe-solution}
```

```
#Use the pipe operator
penguins %>% head()
```

```
#My guess for the last few rows
penguins %>% tail()
```
```

Student practices the described function and then answers a related question.



# Live Demos in Class with R Markdown files



Student submits .rmd file and knitted PDF.  
Instructor can grade from PDF and check  
.rmd file as needed

## The pipe (%>%, read: “and then”)

When writing complex data analysis pipelines, we frequently use the pipe operator %>% to move data from one analysis step to the next. The pipe is pronounced “and then”, and it takes the data on its left and uses it as the first argument for the function on its right.

For example, to see the first few lines of a dataset, we often write `head(data)`. Instead, we can write `data %>% head()`.

Try this yourself. Write code that displays the first few lines of the `penguins` dataset, using %>% and `head()`.

Any guesses on how one can see the last few rows of the data frame?

```
#Use the pipe operator
penguins %>% head()
```

```
A tibble: 6 x 8
species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
<fct> <fct> <dbl> <dbl> <int> <int> <fct>
```

1

```
The pipe (`%>%`, read: “and then”)
```

When writing complex data analysis pipelines, we frequently use the pipe operator `%>%` to analysis step to the next. The pipe is pronounced “and then”, and it takes the data on its first argument for the function on its right.

For example, to see the first few lines of a `dataset`, we often write `head(data)`. Instead, `head()`.

Try this yourself. Write code that displays the first few lines of the `penguins` dataset, `head()`.

Any guesses on how one can see the last few rows of the data frame?

```
```{r pipe-solution}
#Use the pipe operator
penguins %>% head()
```

```
#My guess for the last few rows
penguins %>% tail()
```
```

```
1 Adelie Torge~ 39.1 18.7 181 3750 male
2 Adelie Torge~ 39.5 17.4 186 3800 fema~
3 Adelie Torge~ 40.3 18 195 3250 fema~
4 Adelie Torge~ NA NA NA NA <NA>
5 Adelie Torge~ 36.7 19.3 193 3450 fema~
6 Adelie Torge~ 39.3 20.6 190 3650 male
... with 1 more variable: year <int>
```

```
#My guess for the last few rows
penguins %>% tail()
```



# Live Demos in Class with R Markdown files: uses



To teach R Markdown  
formatting, set up, and options

Guided Exploration  
of help documentation

Learn R Studio

For beginners: have code  
largely written and ask  
students to play  
with arguments

Practice on sample datasets

Practice the code, examples, and  
exercises from course reading

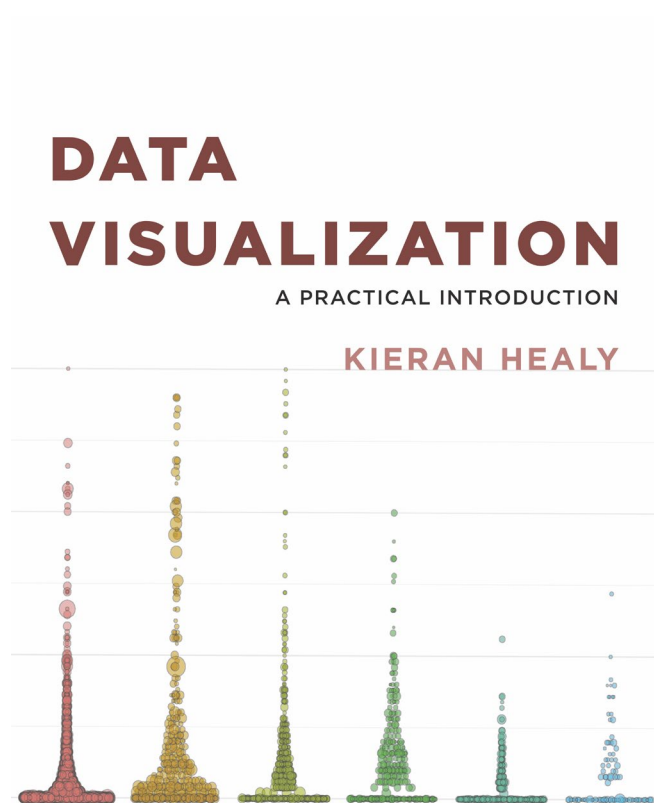
Troubleshoot code, discuss code in paragraph  
form, in addition to commenting



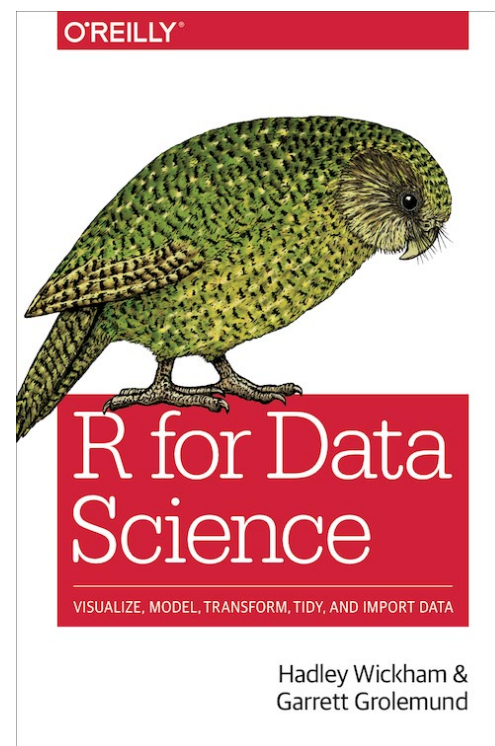
# Live Demos in Class with R Markdown files



- Don't reinvent the wheel! Great markdown files available to accompany lessons in these books:



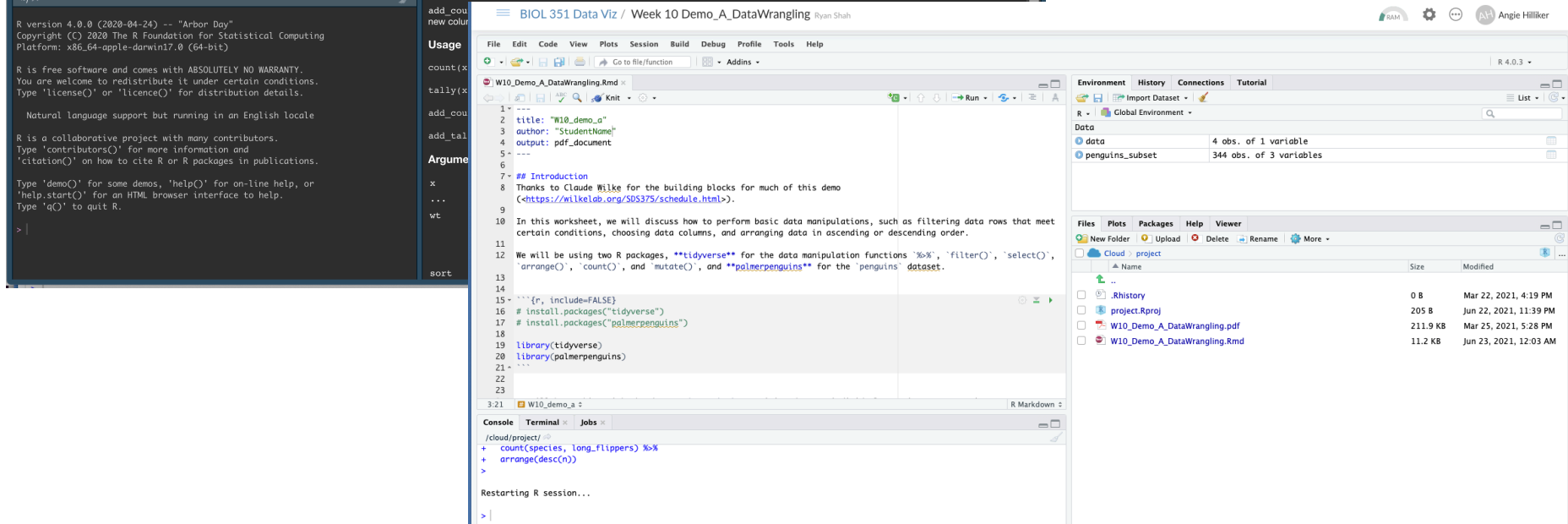
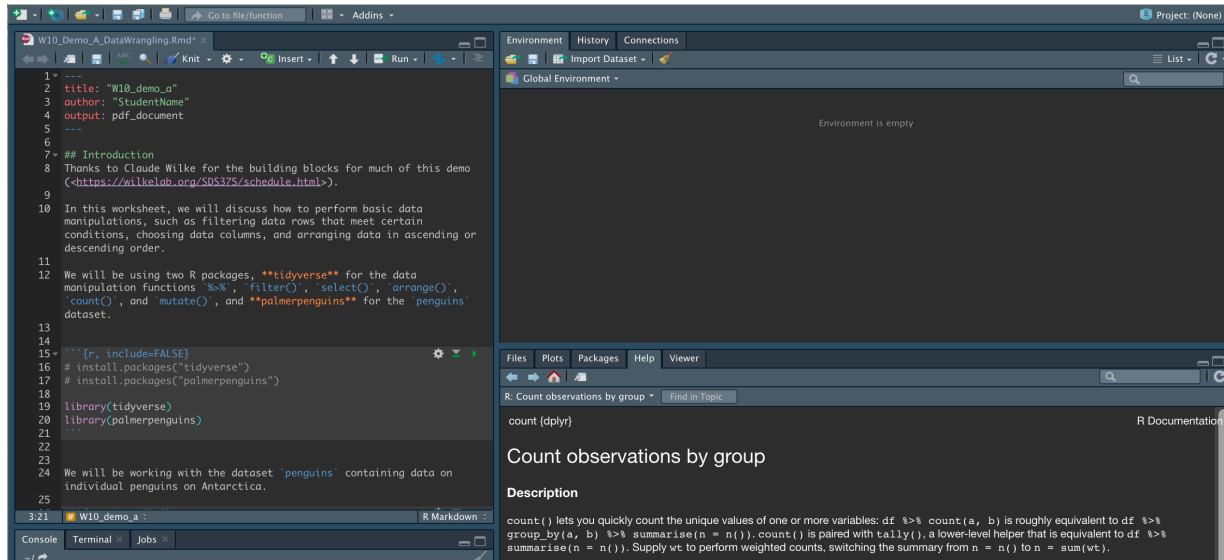
Datasets and .rmd files from [Healy](#)



Datasets and .rmd files from [Wickham](#)



# R Studio Desktop vs. R Studio Cloud



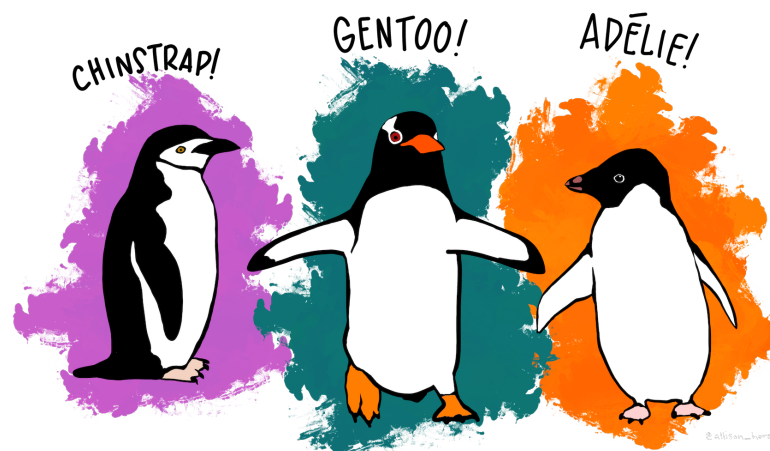
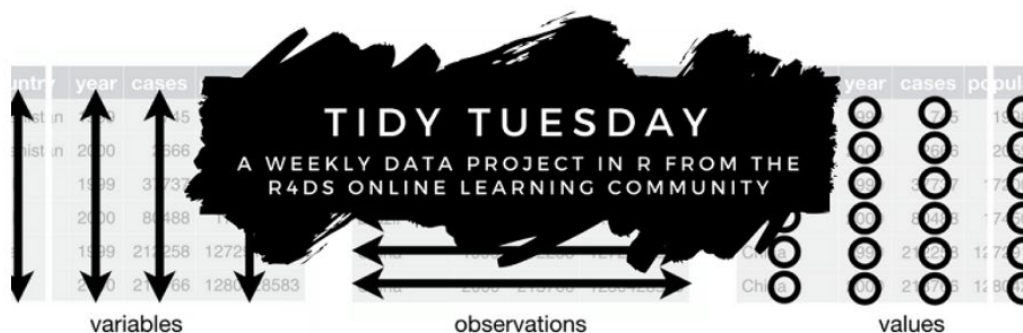


# Tidy Tuesday Challenge



- Weekly social data project to wrangle and visualize data shared via Twitter
- We first asked students to look at past submissions
  - Used [Palmer Penguins](#) week (familiar from tidy data unit)
  - Twitter search or #TidyTuesday.rocks [Shiny app](#)

TidyTuesday - A weekly social data project in R



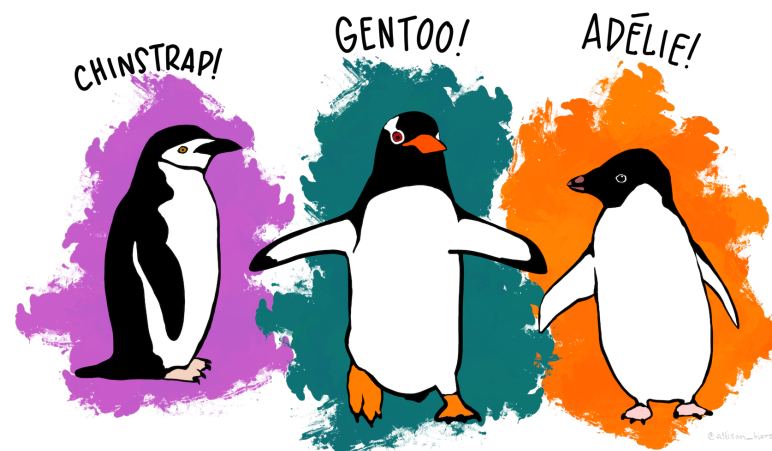
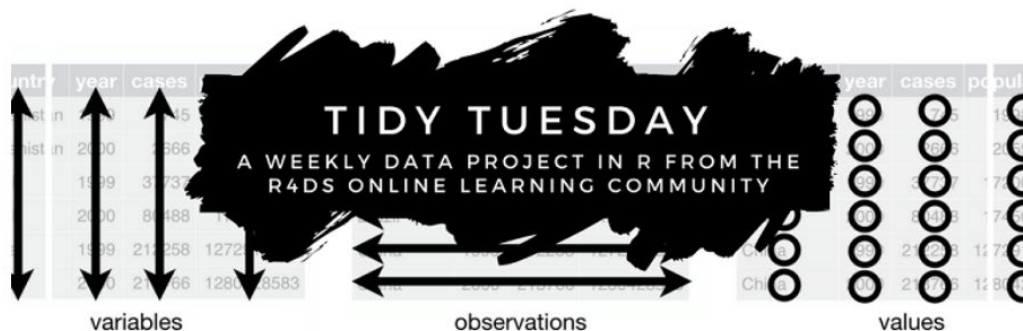


# Tidy Tuesday Challenge



- “When I first imagined a twitter challenge centered around data science, I thought that they would be relatively boring, statistical figures that are more suited for a scientific journal. However, I was clearly incorrect! The submissions that received a lot of likes and retweets tended to be very creative, colorful, and insightful.” – *Student reflection*

TidyTuesday - A weekly social data project in R





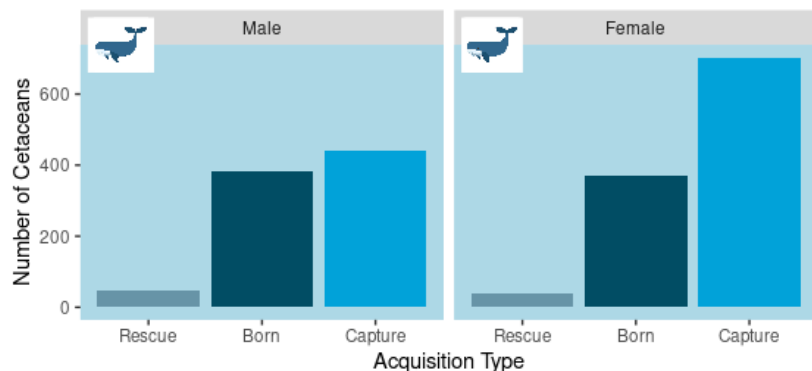
# Tidy Tuesday Challenge



- We selected a [2018 dataset on Cetaceans](#) (whales and dolphins) in captivity in the US for students to select their question and visualize
- Selected for the bio theme and approachability as students built their Program R skills

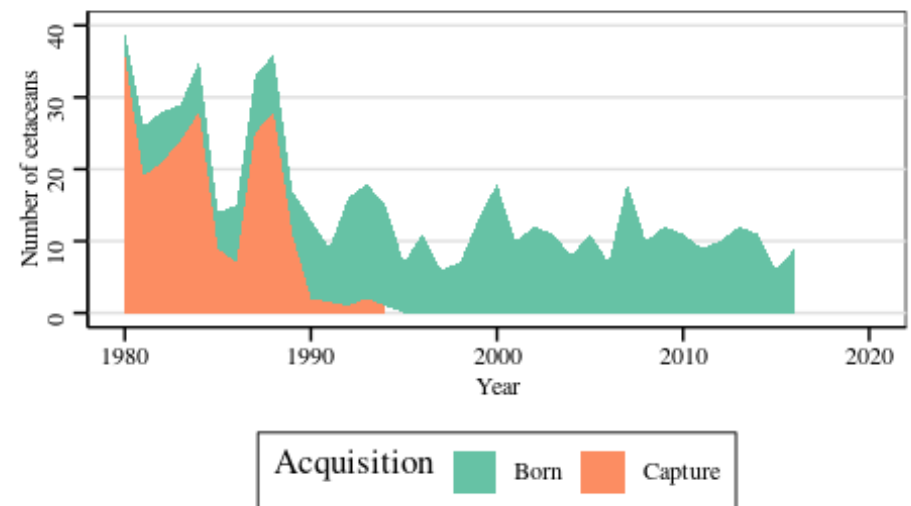
**Number of Rescued, Born, and Captured Cetaceans by Sex in the US From 1938-2017**

More female than male cetaceans have been captured in total



Sources: FOIA, Ceta-Base; collected by Amber Thomas

**How were captive cetaceans acquired over time?**





# **Summative Assignments/Activities**



# Mid-Semester Project



- Partners: Pick a data set from [Our World in Data](#) (OWID)

The screenshot shows the homepage of the 'Our World in Data' website. The header is dark blue with the site's name 'Our World in Data' on the left. In the center, there is a search bar and a navigation menu with links for 'Articles by topic', 'Latest', 'About', and 'Donate'. On the right side of the header, there are logos for 'OXFORD MARTIN SCHOOL', 'UNIVERSITY OF OXFORD', and 'GCDL'. Below the header is a yellow banner with the text 'Coronavirus pandemic: daily updated research and data.' and a 'Read more' button. The main content area has a dark blue background with a faint world map. It features the text 'Research and data to make progress against the world's largest problems' in yellow. Below this is a yellow button that says 'Scroll to all our articles' with a dropdown arrow. At the bottom, it states '3160 charts across 297 topics' and 'All free: open access and open source'.

Our World in Data

Articles by topic

Search...

Latest About Donate

OXFORD MARTIN SCHOOL UNIVERSITY OF OXFORD GCDL

Coronavirus pandemic: daily updated research and data. [Read more](#)

Research and data to make progress against the world's largest problems

[Scroll to all our articles](#) ▼

3160 charts across 297 topics  
All free: open access and open source



# Mid-Semester Project



- Pick a data set from [Our World in Data](https://ourworldindata.org/) (OWID) related to biology





# Mid-Semester Project

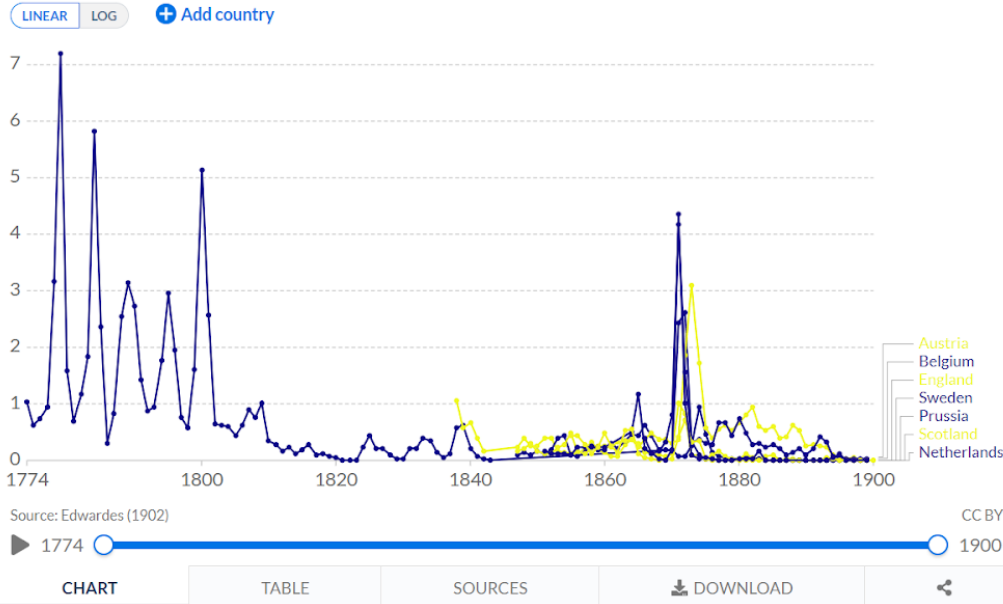


- Each partner selects one Figure from OWID in data to recreate or revise in Tableau

## Deaths from smallpox per 1,000 population

Sweden started vaccinating people against smallpox in 1801. It became compulsory in Sweden in 1816, England in 1853, Scotland in 1864, the Netherlands in 1873, Prussia in 1874, and Belgium and Austria never made the smallpox vaccination mandatory.

Our World  
in Data





# Mid-Semester Project



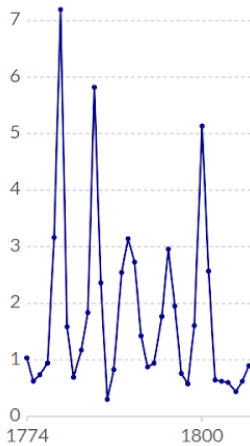
- Each partner selects one Figure from OWID in data to recreate or revise in Tableau

## Deaths from smallpox per 1,000 population

Sweden started vaccinating people against smallpox in 1801. It became compulsory in Sweden in 1816, England in 1853, Scotland in 1864, the Netherlands in 1873, Prussia in 1874, and Belgium and Austria never made the smallpox vaccination mandatory.

Our World  
in Data

LINEAR LOG + Add country



Source: Edwardes (1902)



CHART



## Smallpox deaths per 1,000 population from 1870 to 1880.

Smallpox vaccination became mandatory in Sweden in 1816, and in the Netherlands in 1873. Belgium and Austria never made smallpox vaccination mandatory.

Deaths per 1,000 population





# Mid-Semester Project



- Then explore the data as shown in OWID to find an interesting question or comparison to explore.
- Make a novel figure to explore novel aspect of the data.



# Mid-Semester Project



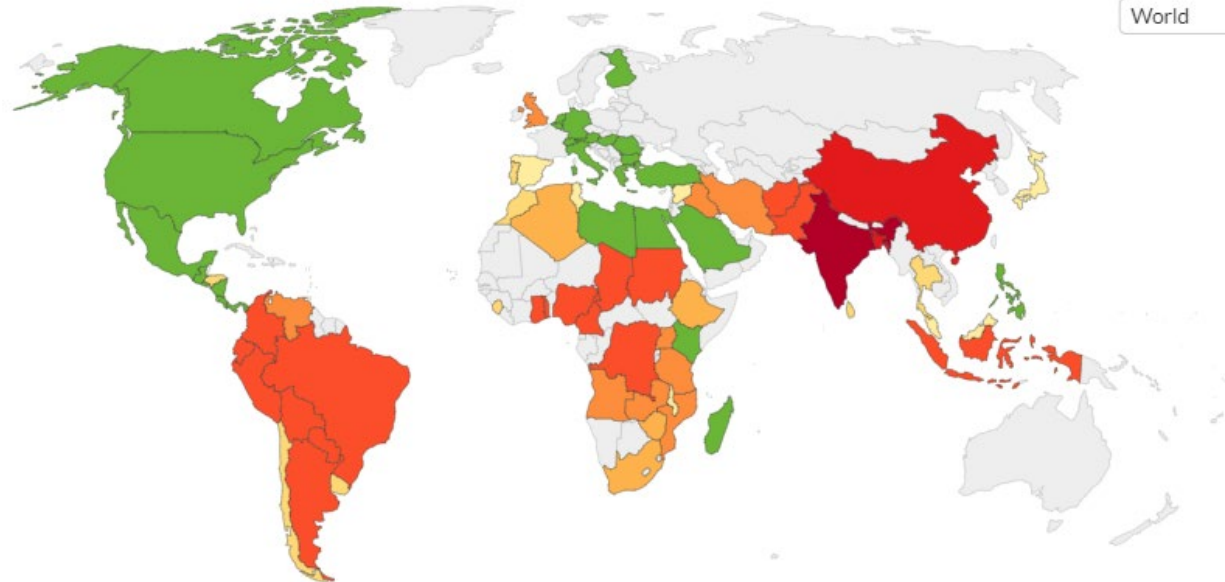
- Make a novel figure to explore novel aspect of the data.

## Number of reported smallpox cases, 1952

This includes endemic and imported cases, therefore some countries still recorded some cases after the official year of eradication.

Our World  
in Data

World



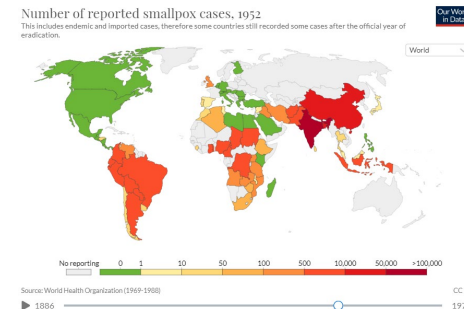
Source: World Health Organization (1969-1988)

CC BY



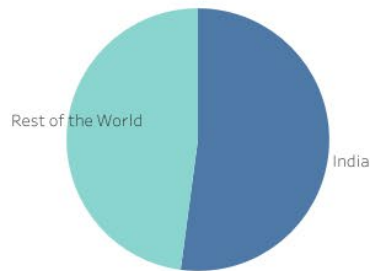


# Mid-Semester Project



## India's Role in High Number of Global Smallpox Cases Between 1925-1977

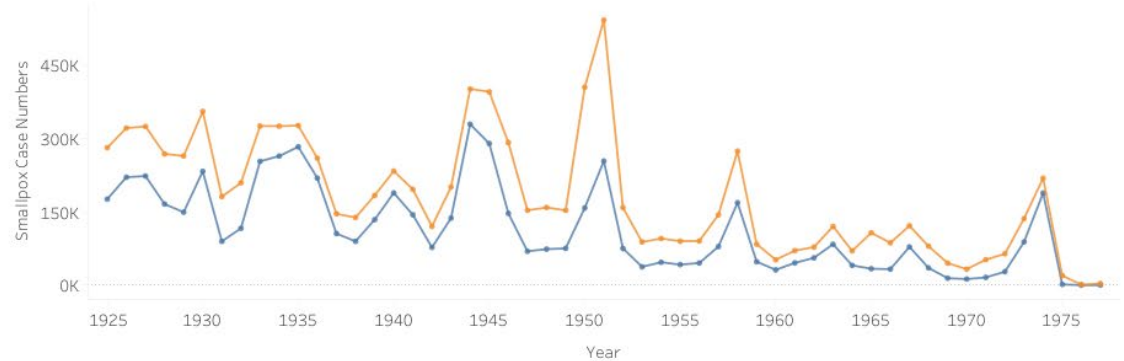
Total Global Smallpox Cases



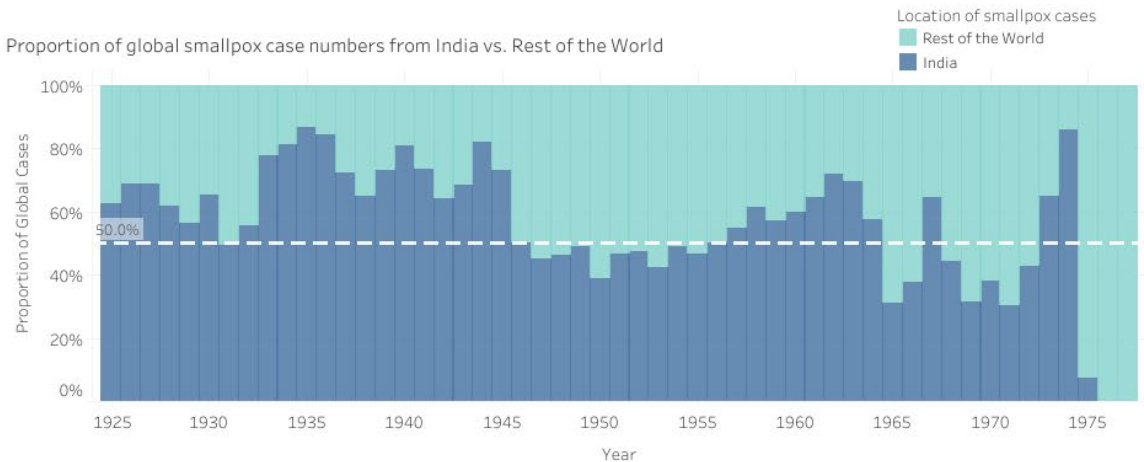
52.11%

of total global cases during 1925-1977  
were from India

Global smallpox cases vs. India's smallpox cases



Proportion of global smallpox case numbers from India vs. Rest of the World





# Mid-Semester Project



- Look in the scientific literature for data related to their topic. Present what you learn from the data and critique the aesthetic choices of the figure.

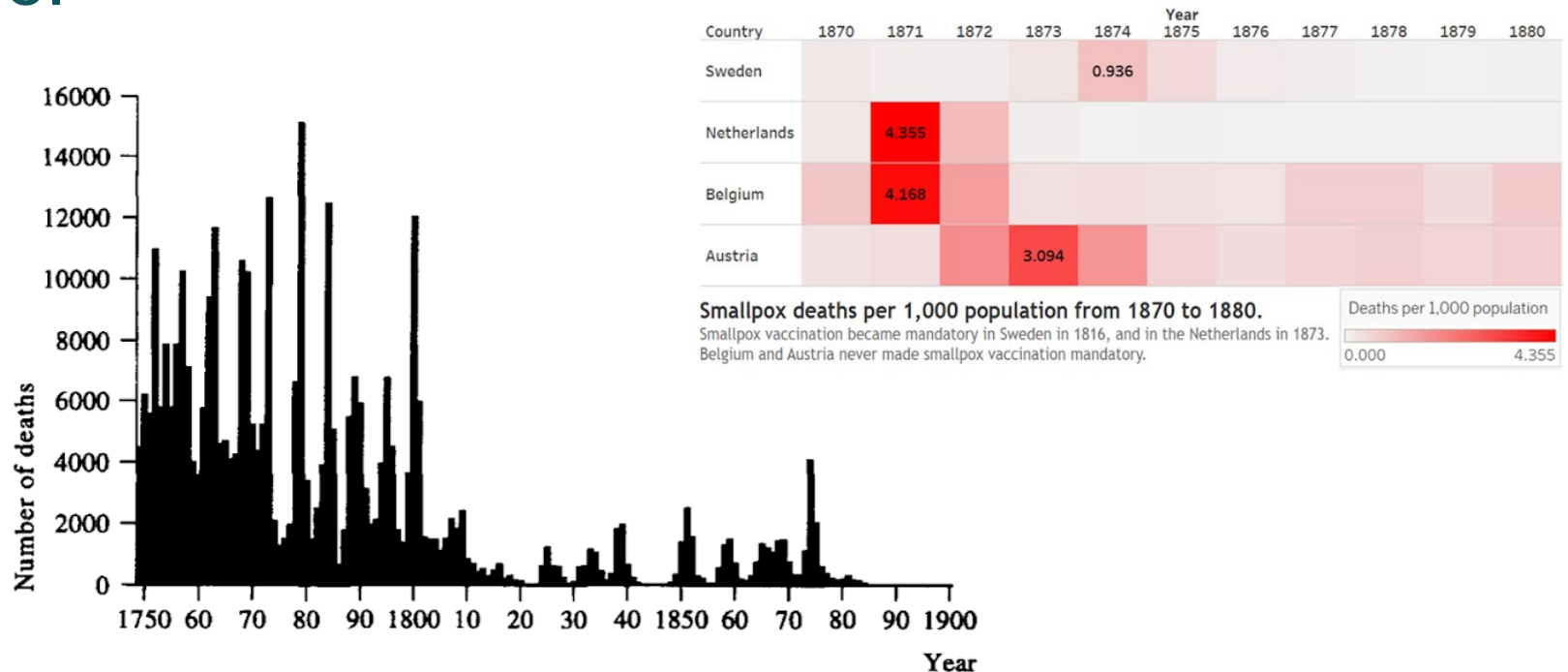


Figure 1. Total number of deaths from smallpox in Sweden 1749–1900. Source: P. Sköld, 'Från fruktad farsot till sällsynt sjukdom. Smittkopporna i Sverige 1750–1900', *Nordisk Medicinhistorisk Årsbok 1994* (Södertälje, 1994), p. 93.



# Mid-Semester Project



- Student deliverables:
  - Oral presentation to class, with slides
  - 2-3 page written analysis
  - PDFs of cited papers
  - Tableau workbook and data



[sciencemag.com](http://sciencemag.com)

## The Weekly Courant, &c.

London, July 5. Bills of Mortality from June 26. to July 3.

|                        |                         |                       |
|------------------------|-------------------------|-----------------------|
| <b>A</b> Bortive ——— 2 | French-Pox ——— 1        | Spotted Fever ——— 5   |
| Aged ——— 17            | Gout ——— 4              | Stillborn ——— 16      |
| Apoplexy ——— 2         | Gripping in the Guts 18 | Stone ——— 1           |
| Childbed ——— 4         | Headmoldshot ——— 1      | Stoppage in Stomach 3 |
| Colick ——— 3           | Jaundies ——— 1          | Stronggullion ——— 1   |
| Consumption ——— 55     | Imposthume ——— 1        | Surfeit ——— 1         |
| Convulsions ——— 113    | Measles ——— 6           | Teeth ——— 15          |
| Dropsie ——— 10         | Mortification ——— 1     | Thrush ——— 3          |
| Evil ——— 1             | Rickets ——— 2           | Tiflick ——— 6         |
| Fever ——— 57           | St. Anthony's Fire — 1  | Ulcers ——— 1          |
| Fistula ——— 1          | Small-Pox ——— 40        |                       |

### C A S U A L T I E S.

Drown'd accidentally 5. Three of them in the River of Thames, viz. One at Albhallows Barkin, One at St. Olave in Southwark, and One at Stains in the County of Middlesex (buried at St. Peters in Cornhill) One in the New River at St. James Clerkenwell, and One St Magnus the Martyr. Kill'd by an accidental Fall from a Window at St. Mary at Rotherith 1. Over laid 1. Shot 2. One with a Pistol, at St James Westminster, and One at St Margaret Westminster. Threw himself out of a Window in St Andrew in Holbourn 1.

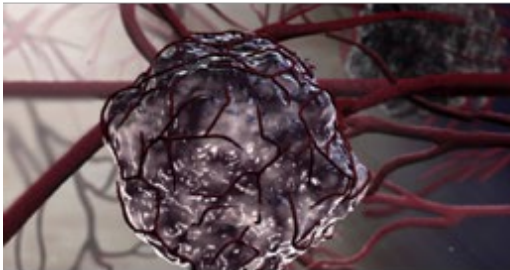
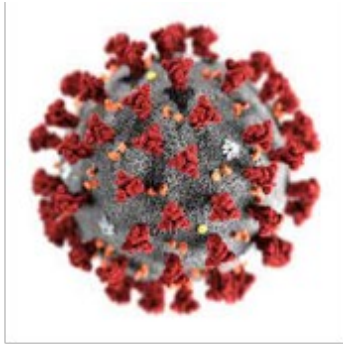
|                                        |      |
|----------------------------------------|------|
| Christen'd ———                         | 326. |
| Buried ———                             | 403. |
| Increas'd in the Burials this Week 21. |      |

Source: [University of Nottingham Blogs](http://University of Nottingham Blogs)



# Capstone Projects

- List of potential topics and data sources to get them started









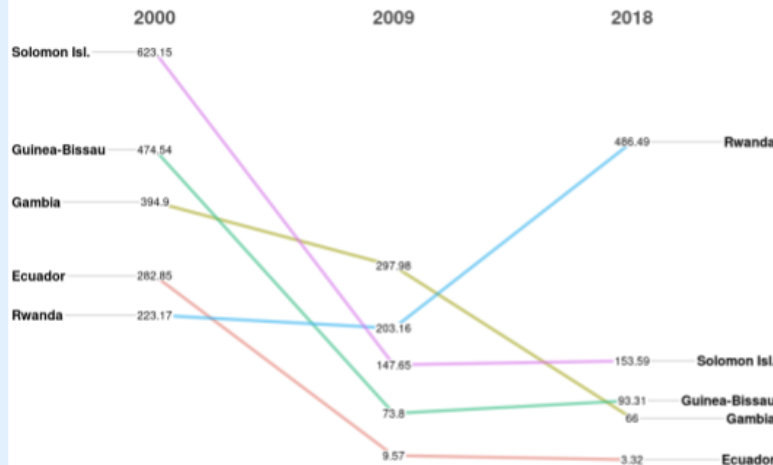
# Capstone Projects

## How have global malaria incidence rates changed over time, and what countries experienced the largest changes?

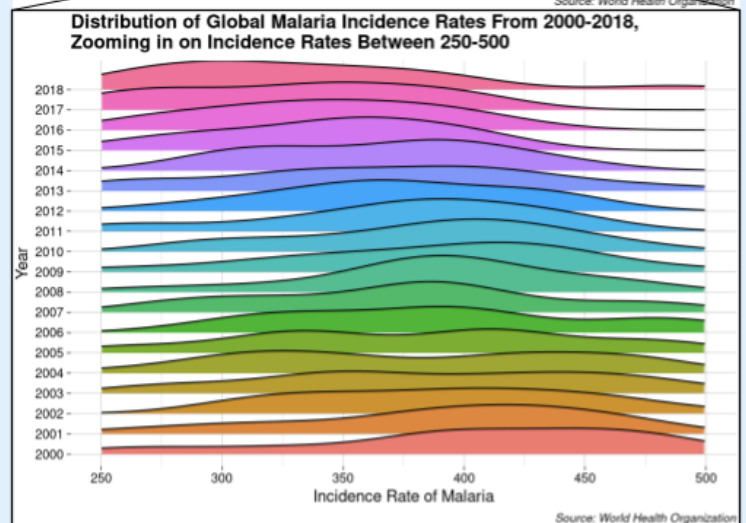
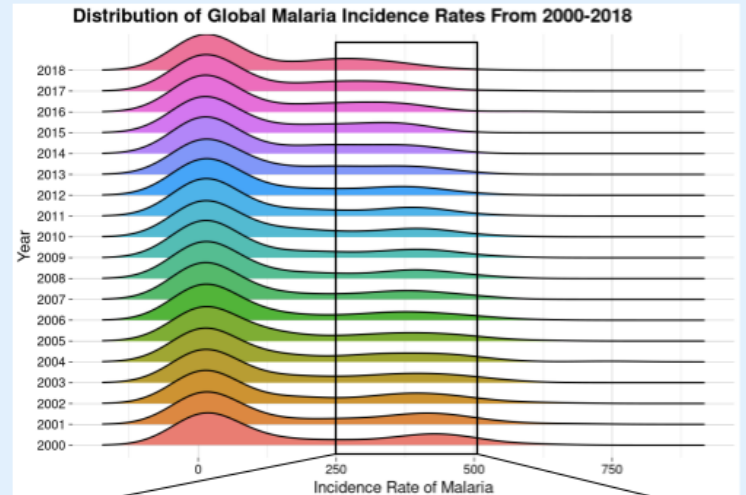
Global incidence rates of malaria have generally declined from 2000-2018, especially for countries with very high incidence rates. This is likely due to increased access to preventative measures such as insecticide-treated nets (Cibulskis *et al.*, 2016).

For the five countries that had the greatest change in incidence rates, four saw dramatic declines from 2000-2018. These five countries were from three different world regions. Only Rwanda experienced an increase in incidence rate, likely due to disruption of public health surveillance systems after the Rwandan genocide (Bizimana & Nduwayezu, 2020).

**Incidence Rates in 2000, 2009, and 2018 for the Five Countries with the Largest Difference Between the Years 2000 and 2018**



Source: World Health Organization



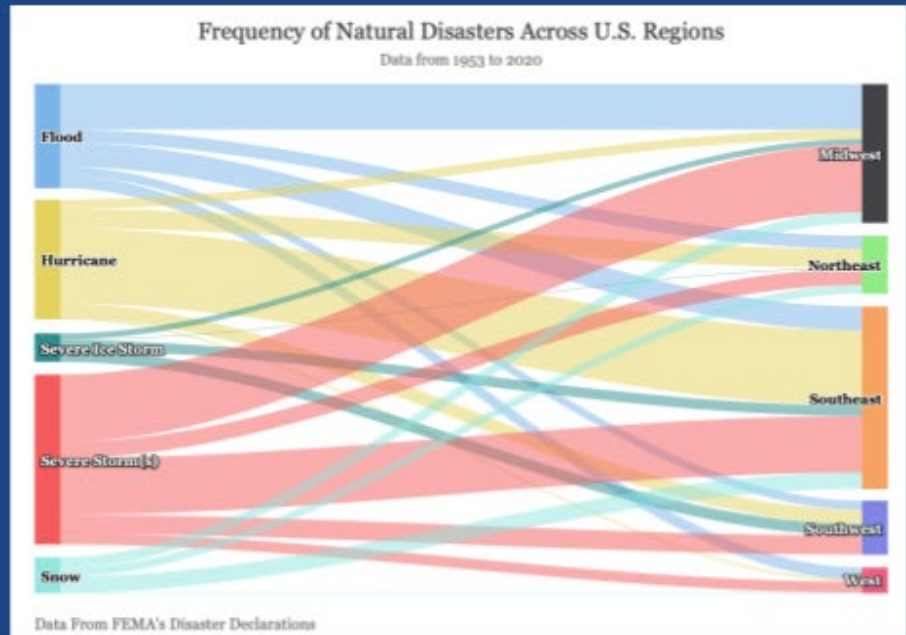
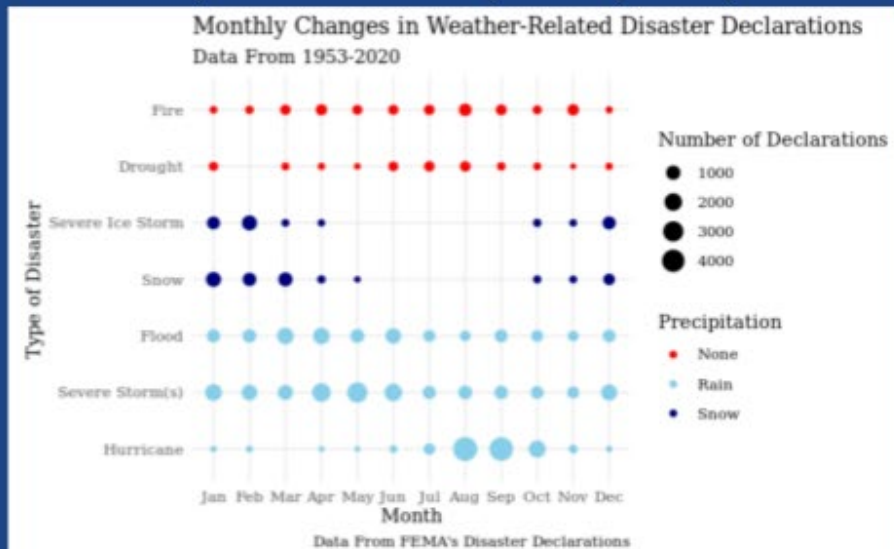


# Capstone Projects

## US Seasonal and Regional Trends in Weather-Related Natural Disasters

Data from FEMA's Disaster Declarations Summary at <https://www.fema.gov/openfema-data-page/disaster-declarations-summaries-v1>

Do types of disasters follow different monthly trends? Are they what you expect?



Are there certain regions of the U.S. that are hit hard by weather-related disasters? Any surprising findings?

Created by Allison Newman



US Seasonal and Regional Trends in Weather-Related Natural Disasters

Background

Natural disasters can influence the way in which we live our lives. Before making decisions about moving to a different state, buying a house, or even traveling to a tropical location, it is essential to consider the risk of hazards and natural disasters. Although geophysical disasters themselves make up roughly half of the natural disaster fatalities, meteorological, hydrological, and climatological events constitute the other half (Hoepppe, 2015). Specifically, precipitation events provide interesting trends. The National Weather Service (NWS) Cooperative Observer Program (COOP), has excelled in tracking these trends and has observed that both frequency and intensity of these precipitation events is increasing nationally in the U.S (Kunkel, 2013). It is hypothesized that this could be a result of increasing water vapor concentrations, since storm precipitation potential is significantly limited by this measure.

As a result, I will use this project to explore trends associated with climate and precipitation-related natural disasters. Weather can be unpredictable, but can we predict which regions will be affected and when? Questions like these are crucial in thinking about mitigating natural hazards and protecting those that live in these locations. There are many programs, such as the Hazard Mitigation Program, that receive money to work with public officials to respond to hazards and reduce risks of future destruction by similar events (Disaster declarations summaries -v2). Therefore it is crucial to use spatial and temporal trends of natural disasters to think about what specific projects should have priority.

Data Overview

The data set explored is a csv file (DisasterDeclarationsSummaries.csv) downloaded from the Federal Emergency Management Agency's website (Disaster declarations summaries -v2). This raw data set covers disaster declarations that were federally declared in the U.S. from 1953 to 2020. A disaster is federally declared when assistance from the federal government is needed or a disaster is deemed an emergency. Data was officially recorded by FEMA when National Emergency Management Information System (NEMIS) was introduced in 1998. This data contains 61, 667 observations and 19 variables. The variables that I am considering include incident type, state, and begin date of the incident. Note that each row represents a disaster declaration, and not a single disaster. Thus multiple rows could constitute different declarations for a single disaster.

Data Visualization

For my first figure, I want to explore the seasonal trends of different disaster incident types that are related to the weather. Furthermore, I would like to break down the trends by type of precipitation: no precipitation (i. e. drought, fire), snow, and rain. I hypothesize that within each precipitation type, there are visual seasonal trends. I also predict that floods will correspond with higher numbers of snow and rain precipitation because snow ultimately melts to water.

For my second figure, I want to explore the regional trends of different disaster incidents that are most closely associated with weather. Thus, I don't consider drought and forest fires in order to cut down on the number of nodes in the sankey chart to make it more manageable. I believe that while some disaster types will appear to be evenly distributed across the U.S., other disaster types will occur more frequently in specific regions. For example, I predict that the northeast will experience the most snowstorms and ice storms, and the southeast will see the majority of hurricanes.

```
#To install the necessary packages to be used throughout this project
#install.packages("dplyr")
library(dplyr)

##
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':
##
filter, lag

The following objects are masked from 'package:base':
##
intersect, setdiff, setequal, union

#install.packages("tidyverse")
library(tidyverse)
```

```
#To create a data set that extracts the months from the date
monthly_trends <- disasters %>%
 #Change data type to 'date' by removing unnecessary characters
 mutate(date = as.Date(str_remove_all(incidentBeginDate, "[00:00:00.000]"))) %>%
 #Extract month
 mutate(month = month(date)) %>%
 #Extract year
 mutate(year = year(date)) %>%
 #Summarize data and create a dot plot
 monthly_trends %>%
 #Choose precipitation-related incidents
 filter(incidentType != "Severe Storms", "Flood", "Severe Ice Storm", "Snow", "Drought", "Hurricane") %>%
 mutate(precipitation = ifelse(incidentType %in% c("Severe Storms", "Flood", "Hurricane"), "Rain",
 ifelse(incidentType %in% c("Snow", "Severe Ice Storm"), "Snow", "None")) %>%
 group_by(month_name, incidentType, precipitation) %>%
 #Summarize number of incidents/month
 summarize(count = n()) %>%
 #Order the levels
 explicit(levels = factor(month_name, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))) %>%
 #Include position or indication of number of incidents and also precipitation type
 geom_point(aes(x = count, color = precipitation)) +
 scale_color_manual(values=c("red", "skyblue", "navyblue")) +
 theme_minimal() +
 #First theme with an interesting font
 theme(text = element_text(family = "Georgia")) +
 #Choose labels
 labs(
 title = "Monthly Changes in Weather-Related Disaster Declarations",
 subtitle = "Data From 1953-2020",
 caption = "Data From FEMA's Disaster Declarations",
 color = "Precipitation",
 size = "Number of Declarations",
 x = "Month",
 y = "Type of Disaster"
)

'summarize()' has grouped output by 'month_name', 'incidentType'. You can override using the '.groups' argument.
```

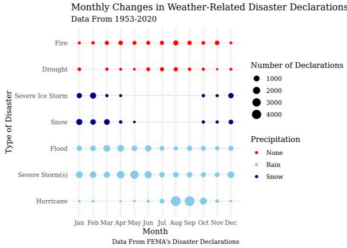


FIGURE 1. Dot plot of monthly changes in weather-related natural disaster declarations in the U.S. Dots are colored by precipitation type, while size is represented by the number of declarations made by FEMA. Droughts and fires are included because of their association with a dry and climate.

Data Analysis

According to Figure 1, fire and drought natural disaster declarations appear to be fairly in sync but there isn't a significant monthly trend. However, the month of August happens to be a more eventful month for fires and droughts. There are also clear trends between snow and severe ice storms, with more snow disasters than severe ice storms throughout the early spring months of March, April, and May. Floods and severe storm monthly trends are very similar, with hurricane disasters deviate from the other rain-associated disasters. Hurricanes are most frequent during the months of August and September, which corresponds to hurricane season. Figure 2 displays significant trends in terms of uneven distributions of natural disasters across the regions. For example, the southeast and midwest receive a majority of the severe storms, while floods mostly affect the midwest. I was surprised by the low number of snow and severe ice storm disaster declarations in the northeast. The southeast actually experiences a great amount of snow storm disaster declarations even though I assumed that the northeast was hit hardest by these storms. Overall, I learned that fires and droughts do not follow very clear monthly trends like disasters that result in precipitation. Perhaps, droughts and fires aren't as highly correlated with weather and climate as I originally presumed. In addition, the southeast and midwest receive the greatest proportion of natural disasters. Surprisingly, the southeast experiences many snow and ice disaster declarations despite their warmer weather.

The strength of these results is supported by the sample size of the data set. The data includes a very large number of disaster declarations and it covers a wide range of years. However, there does appear to be a lack of data from years prior to 1990, which is most likely a result of retrospective data collection done prior to that point. In addition, I explored the number of disaster declarations, where a single disaster could have multiple declarations based on the number of areas affected. While this variable is great for comparing the number of disasters and the counties affected, it should not be confused with the number of unique disasters. In the future, I would also explore the fatalities or losses associated with these disasters. Unfortunately, this specific data is not capable of exploring that question. It would also be interesting to identify the seasonal trends in geological natural disasters, including volcanic eruptions and earthquakes. I would expect these natural disasters to be prone to stochastic effects and random chance, but this hypothesis should be confirmed.

My visualization choices aid in effectively communicating the messages I discussed. The size and color differences in Figure 1 allow for the audience to make conclusions about frequency trends between different precipitation types and the monthly changes. The simple white background with light gray gridlines is needed to allow the dots to stand out and connect these dots to their respective months and disaster types. In Figure 2, the consistency in color choice between the sankey bars and the disaster type help to trace the flow back to its source. In addition, there was not chosen for this plot because none of the themes come with enough colors to differentiate all of the disaster types. In the future, I would determine if it is possible to create my own colors for each source and bar. After thorough research, I did not find usable code.

Discussion

Although I was surprised by some of my findings, the literature does confirm the accuracy of my results. For example, the number of snowstorms in the southeast should be considered in the context of temporal trends. For example, one paper reports that the number of severe snowstorms in the southeast U.S. has drastically declined when comparing data after and before 1980 (Gutierrez & LePewest, 2016). This trend is concerning because these storms can result in damages such as power outages, as we have seen this year in Texas. Furthermore, even as the snow seasons become warmer and dryer, severe snow and ice storms are continuing to occur (Kunkel et al., 2013). It would be interesting to align this data with climate change data to examine any correlations in severe, warm snowstorms.

In terms of fires, forest fires are associated with spring and summer seasons (Westering, 2016). However, this trend is reliant on "warm spring and summer temperatures and early spring snowmelt dates" (Westering, 2016). In fact, Westering found that the length of the fire seasons from 2003-2012 were 84 days longer than the fire seasons in 1973-1982. While this reported range does not explain why Figure 1 displays fires throughout the year, this statistic might contribute to this trend. In addition, the "fire" category in Figure 1 consists of forest fires and other possible fire types that could be declared emergencies. This is important to consider when interpreting the lack of seasonal changes in frequency.

While my results confirm expected seasonal and regional differences in severe weather-related natural disasters, the literature is able to support some of the major deviations from my results and my hypothesis. From this data exploration, we can better utilize the frequency of weather-related disasters declared to open mitigation projects to protect these regions during vulnerable times. Lastly, this project opens up many questions about climate change, which could be visualized using yearly trends.

References

Barnes, B., Dunn, S., & Wilkinson, S. Natural hazards, disaster management and simulation: a bibliometric analysis of keyword searches. *Nat Hazards* 97, 813–840 (2019). <https://doi.org/10.1007/s11069-019-03677-2>

Disaster declarations summaries -v2. (n.d.). Retrieved April 30, 2021, from <https://www.fema.gov/openfile-data-page/disaster-declarations-summaries-v2>

Hoepppe, Peter. (2015). Trends in weather related disasters – Consequences for insurers and society. *Weather and Climate Extremes*. 11, 10-10(6) <https://doi.org/10.1016/j.wace.2015.10.002>

Gutierrez, K. S., & LePewest, C. E. (2016). *Climate Justice in Rural Southeastern United States: A Review of Climate Change Impacts and Effects on Human Health*. International journal of environmental research and public health, 13(2), 189. <https://doi.org/10.3390/ijerph13020189>

Kunkel, K., Karl, T., Brooks, H., Kossin, J., Lawrence, J., Arndt, D., ... Wuebbles, D. (2013, April 01). Monitoring and understanding trends in extreme storms. *State of knowledge. Bulletin of the American Meteorological Society*, 94(4), 499–514 <https://doi.org/10.1175/BAMS-D-11-00062.1>

Pawar, V., & Sen, S. (2019). Economic Impact of Natural Disasters: An Empirical Re-examination. *Margin: The Journal of Applied Economic Research*, 13(1), 109–139. <https://doi.org/10.1177/0973801018800807>

Smith, A. B. (2020). U.S. billion-dollar weather and climate disasters in historical context. *NOAA Climate.gov*. (2021, January 08). Retrieved April 30, 2021, from <https://www.climate.gov/news-features/infographic/billion-dollar-weather-and-climate-disasters-historical>

Westering, A. L. (2016). Increasing western US forest wildfire activity: sensitivity to changes in the timing of spring. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 371(1046), 201501176. <https://doi.org/10.1098/rstb.2015.01176>



**What we learned:**



# What we learned

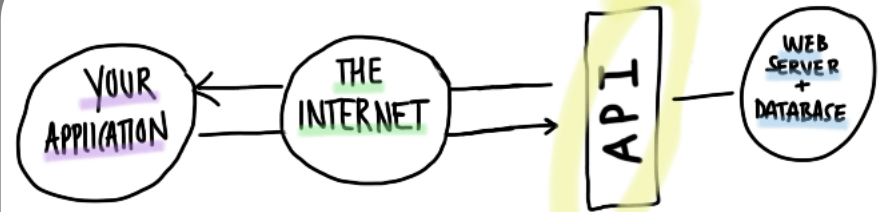
- Many great open source books with exercises
- Instructors share activities and data sets to work with (their websites, Twitter, QUBES, etc.)
- Do consider that “keys” and sample answers may be online for open source texts. How do you want to handle that?
- Blackboard as a landing pad for each class with links to various resources
- Google slide deck to create individual “Data Viz Journal” for turning in small assignments



# We couldn't include several fundamental coding concepts:

```
1 say_hello <- function(name) {
2 message <- paste("Hello", name)
3 return(message)
4 }
```

[datascibc.org](https://datascibc.org)



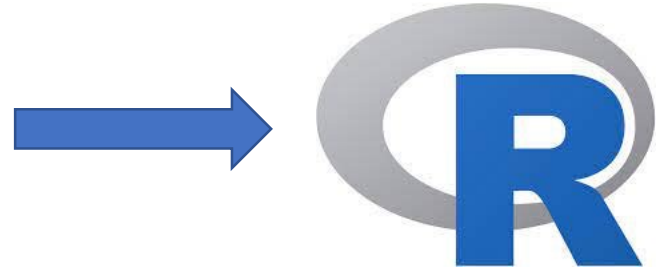
[Elle Townsend](#)



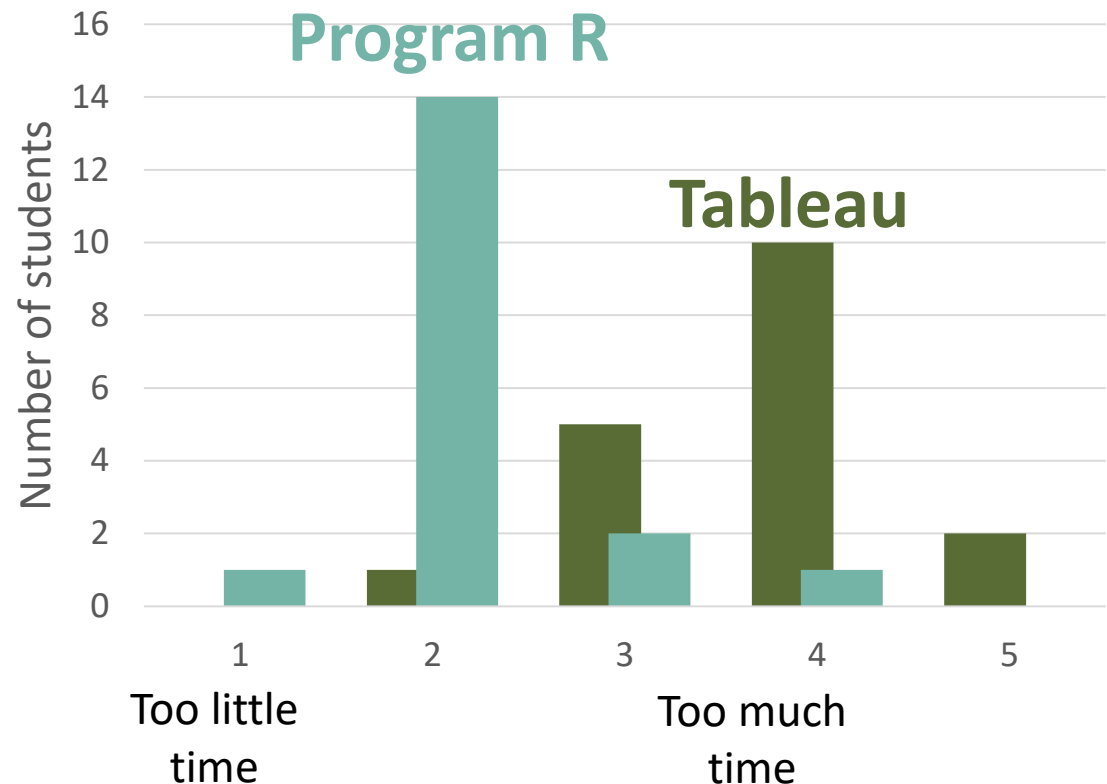
[promptcloud.com](https://promptcloud.com)



# Changes we will make: focus on one program so we can introduce more skills

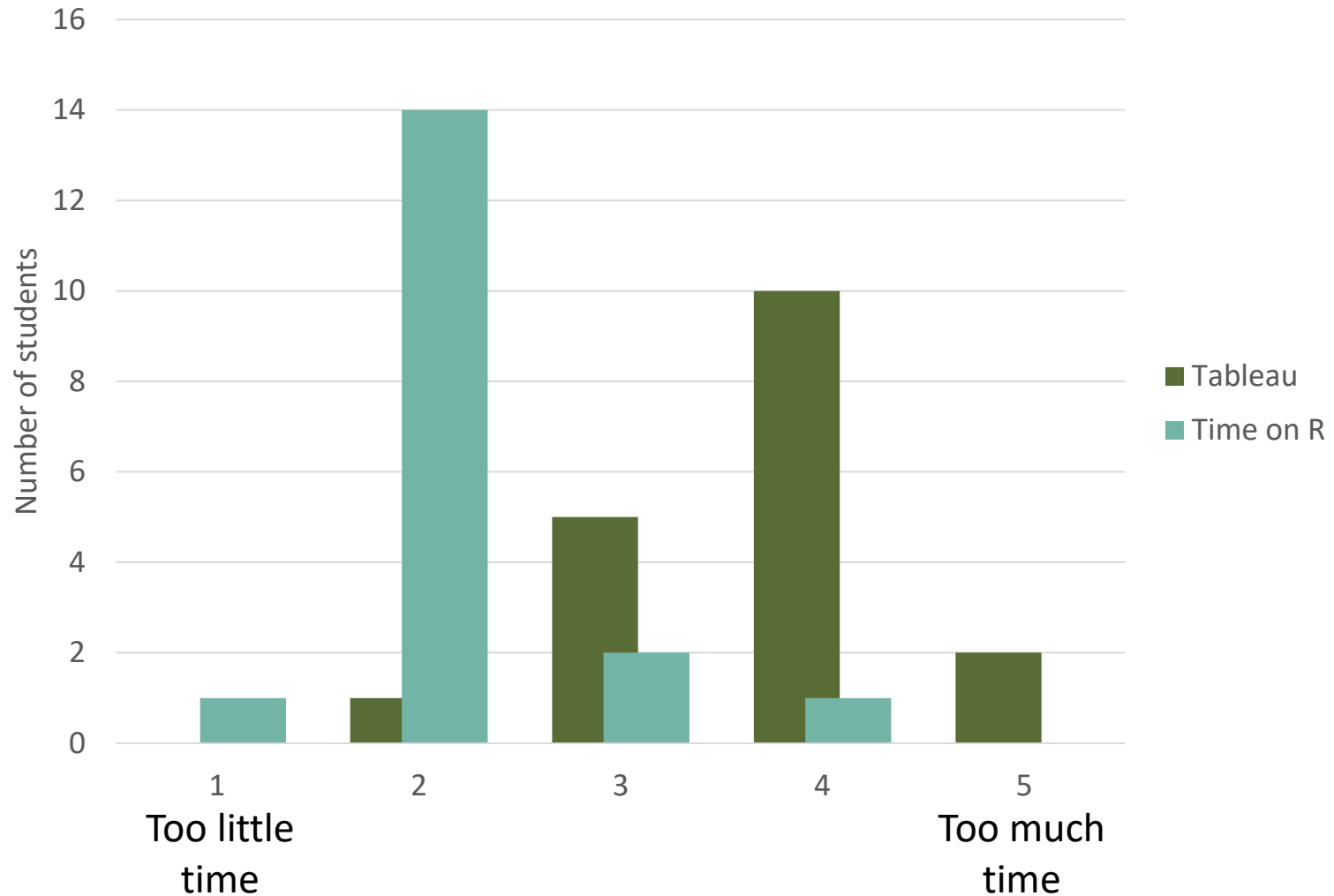


**Student survey:  
Rate the amount  
of time spent on  
Tableau and R:**





# Student survey: Rate the amount of time we spent on Tableau/ Program R:

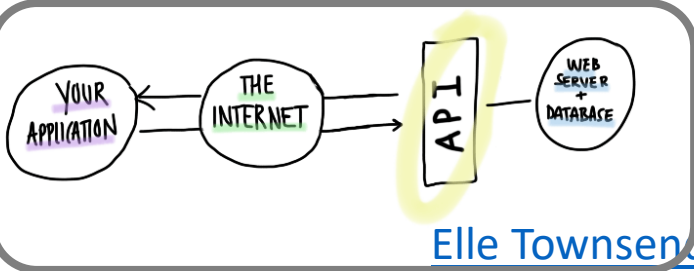




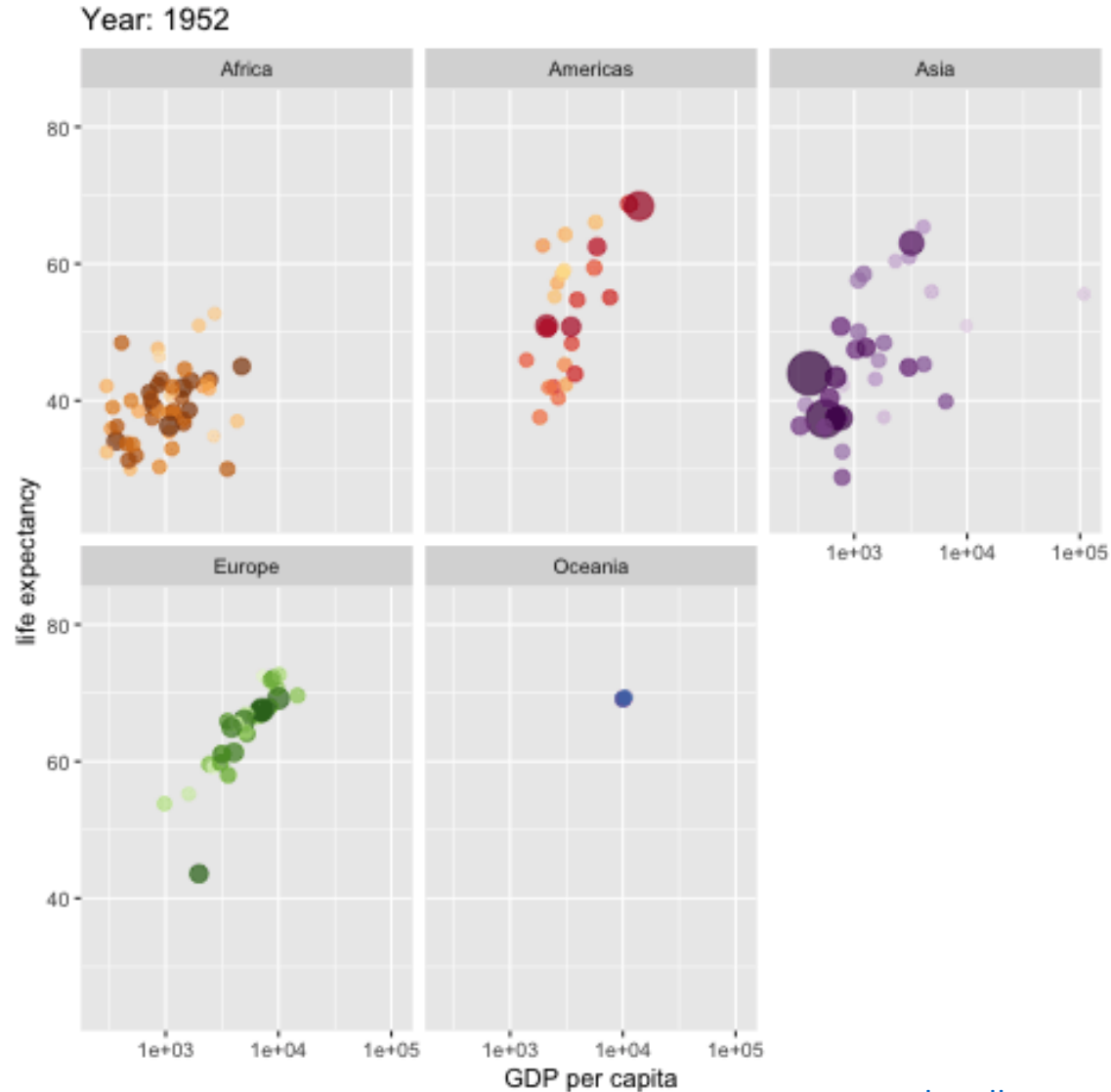
# What would we add with more time?

```
1 say_hello <- function(name) {
2 message <- paste("Hello", name)
3 return(message)
4 }
```

[datascibc.org](http://datascibc.org)



Elle Townsend



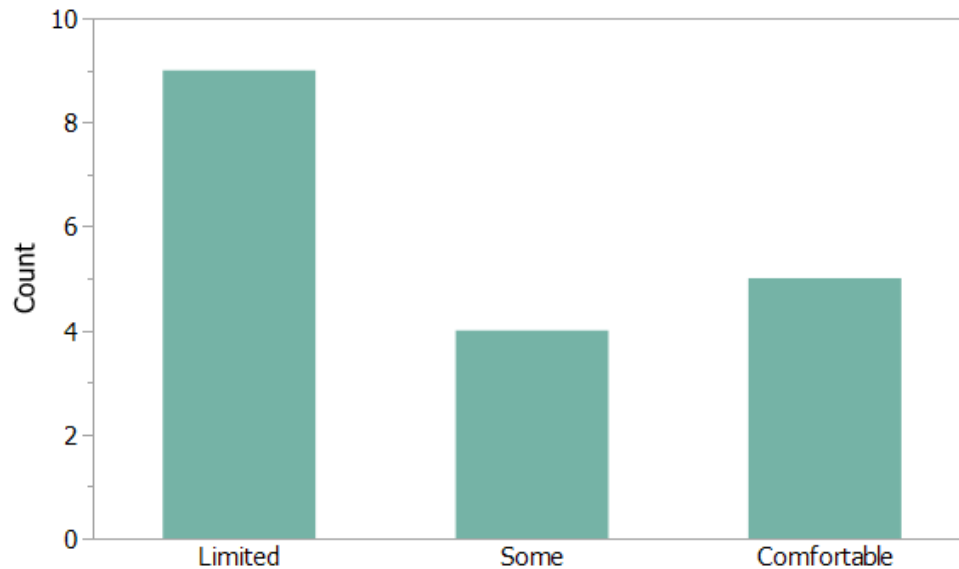


# **Student Feedback:**



# How would you characterize your comfort and skills for working with data?

Before starting this class:



- “I was pretty unfamiliar with even the most basic principles of data management and visualization. My only real experience involved basic R from BIOL202 & whatever Excel skills I have picked up over the years.”



# Think about where you are now. What has changed and what have you learned that you value?

- “I actually know how to approach large datasets now, which is exciting. I also feel more confident in my ability to create appealing and effective visualizations. I especially value the skills that I have learned for Tableau and R, since I can see myself using these tools to some capacity in the future.”
- “This class pushed me out of my comfort zone and has provided me such a better understanding of various data visualization methods. I've begun to see data visualization as a much more intricate process, whereas before I would go through a similar pattern of motions when making graphs on excel.”



# Thanks to:

- All materials co-created with Dr. Kristine Grayson
- Thank you to Krista Stenger (Chair of Biology) for supporting the vision of this course
- Thank you to Andrew Bell and Linda Boland (UR's Faculty Hub) for supporting the expansion of Data Literacy curriculum at UR
- Great online resources used from:
  - Carrie Eaton
  - Claude Wilke
  - Andrew Heiss
- Our hard working students!

