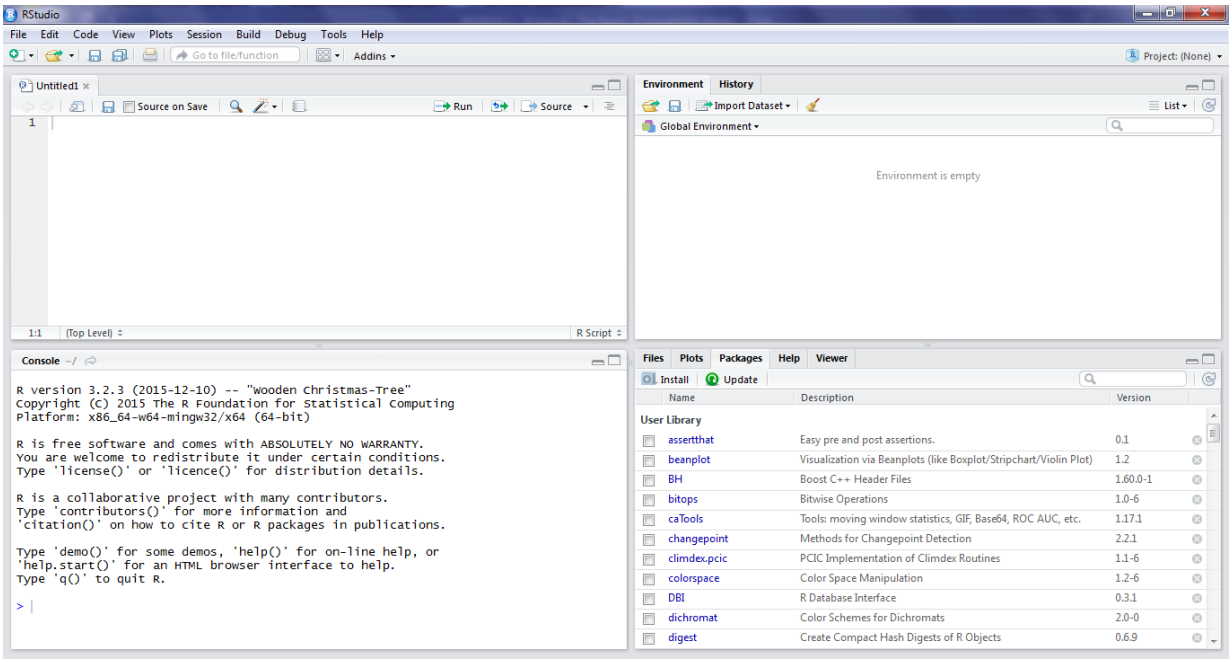
Instructions to Download R, RStudio, and Packages

A) You must download base R first, since it must be installed on your computer for RStudio to work.
   1) Go to this webpage: https://cloud.r-project.org/
   2) Under "Download and Install R," click the link that corresponds to your computer's operating system.
   3) FOR WINDOWS:
      a) In the first line (after "base"), click the link "install R for the first time".
      b) Click "Download R 4 …" at the top of the page.
   4) FOR MAC:
      a) Under "Latest Release", click on the first file name that ends in ".pkg".
   5) After downloading, proceed as you normally would to complete installation of a program.

B) Next, install RStudio – this requires base R to already be installed on your computer.
   1) Go to this webpage: https://www.rstudio.com/products/rstudio/download/#download
   2) Under "All Installers," click the link that corresponds to your computer's operating system.
   3) After downloading, proceed as you normally would to complete installation of a program.

C) I recommend opening RStudio after installation is complete to ensure it runs correctly. You should see a screen that looks something like this:



D) Install four packages in RStudio that we will use during the workshop.
   1) Open RStudio.
   2) Click the "Tools" toolbar at the top of the screen.

3) Click “Install Packages…”.
4) Make sure “Repository (CRAN)” is selected in the “Install from:” box.
5) Type the names of packages to install in the “Packages” box:
   `ggplot2, ggpubr`
6) Make sure “Install dependencies” is checked.
7) Click “Install”.
8) If successfully installed, you should see something like this in the “Console” panel:

If you have any problems during the installation,
please email me (pillarm@miamioh.edu) and I’ll do my best to help you out!

Computer Science in Modern Biology

Data Visualization in R

Instructor:
Rachel Pilla

TAs:
Mason Murphy

Mahbobeh Lesani

1

Workshop Schedule

Day 1 (Wednesday):

- Structure and design of graphs
- Creating graphs in base R
- Overview of ggplot2

Day 2 (Thursday):

- Introduction to ggplot2
- Building graphs and customizing in ggplot2
- Additional graphing resources
- Q&A and practice

2

Workshop Logistics

Zoom
- Feel free to unmute yourself to ask questions at any point

Chat
- Post any questions or comments here
- The TAs will be checking the chat to help answer questions

Break-Out Rooms
- If you are really stuck, you can move to a break-out room with a TA to share your screen and get you up to speed
- Break time is a great opportunity for this if needed!

3

Packages in R

- ***Packages*** are bundles of tools and functions that others have developed to be used in R
- They are often grouped to specific types of functions, analyses, or datasets
- `rLakeAnalyzer`, for example, has lots of functions developed by limnologists to help you analyze common types of data collected from lakes

4

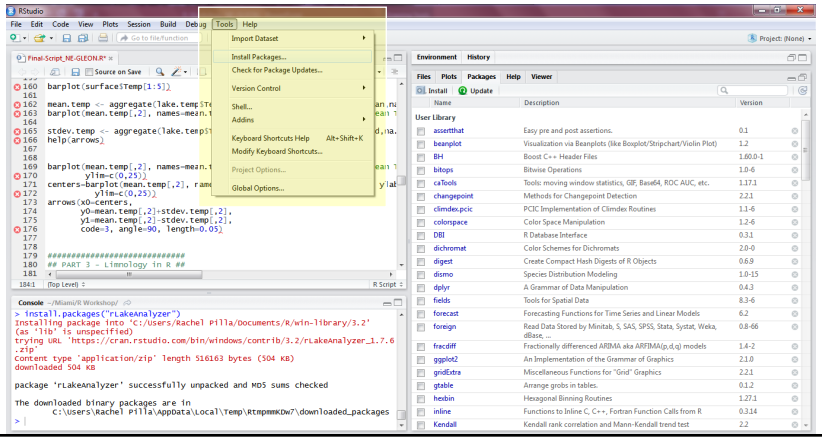
Packages in R

- Currently, there are over 16,000 packages!
- We are going to install 4 packages to use today:
  1) `ggplot2`
  2) `ggpubr`

5

Installing a Package

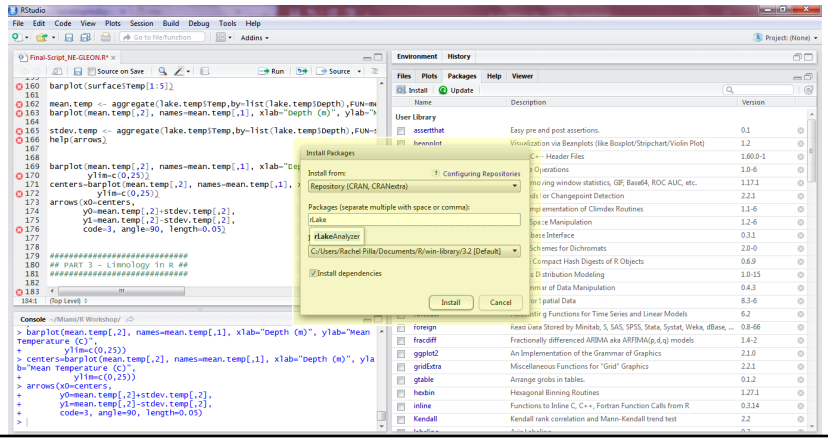
- Tools → Install Packages…



6

Installing a Package

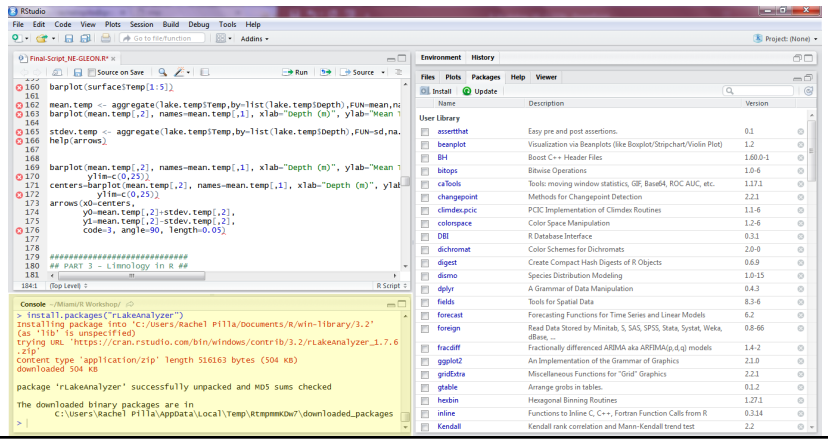
- Type the packages you want to install:
  `ggplot2, ggpubr`



7

Installing a Package

- Console will tell you when they have been installed



8

Installing a Package

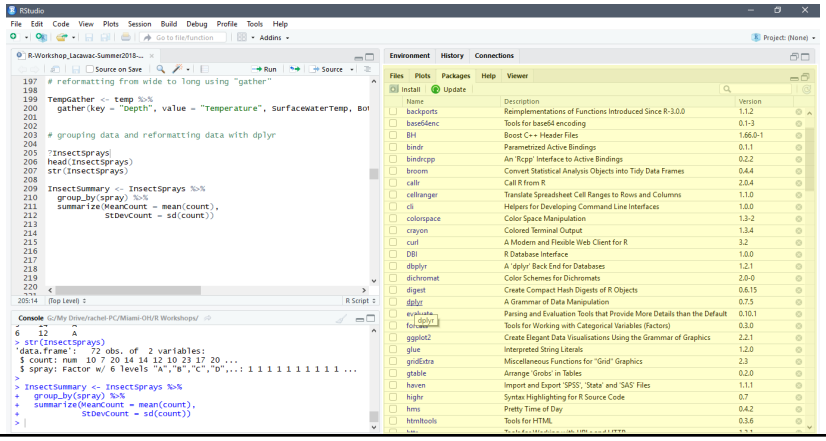
- Packages only need to be installed ONCE
- But, each time you re-open R, you need to load the package(s) you want to use so R can make all the functions available:

```
library(package name)
```

9

Package Information

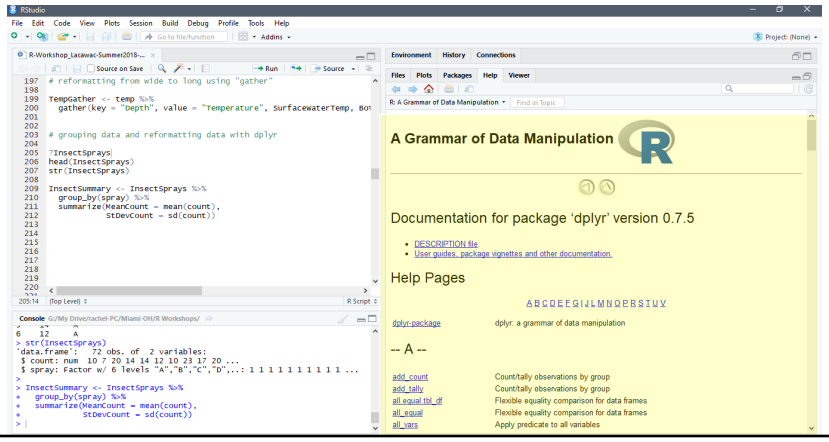
- Under the "Packages" tab, click on **`dplyr`**



10

Package Information

- This will list all the available functions
- Clicking on a function takes you to the help file



11

Structure & Design of Graphs

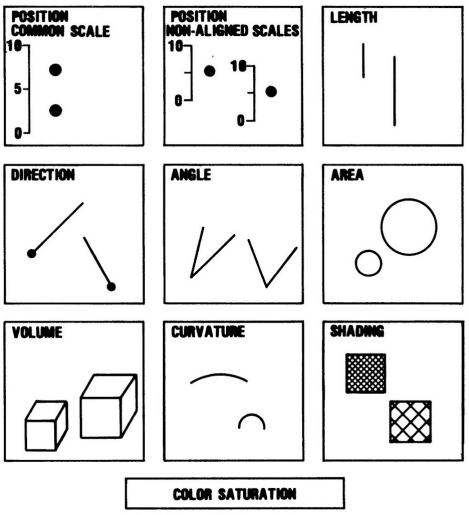
12

Components of a Graph

1) Axes
   - Horizontal x-axis and vertical y-axis
2) Labels
   - What are your axes measuring?
3) Title and/or caption
   - Description and information about the graph
4) Legend (sometimes)
   - If multiple lines, colors, shapes, etc., what does each represent?
5) DATA
   - The actual information!

13

Understanding a Graph



- Easy to perceive and understand

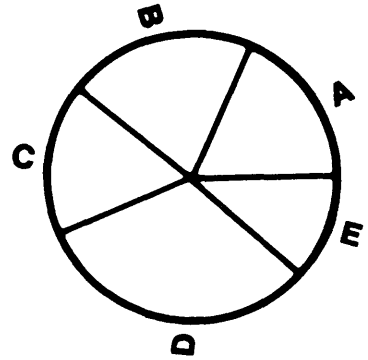
↓

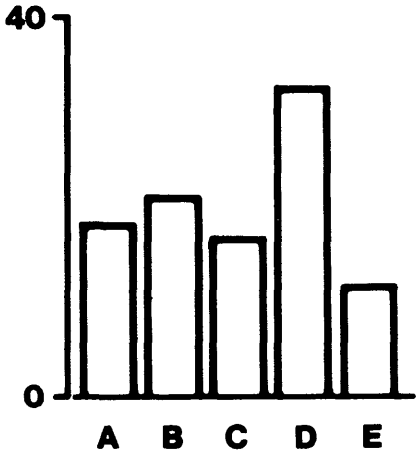
- May require more work to understand
- Can be very valuable when used appropriately in graphs

Cleveland & McGill, 1984 (*Journal of the American Statistical Association*)

14

Understanding a Graph



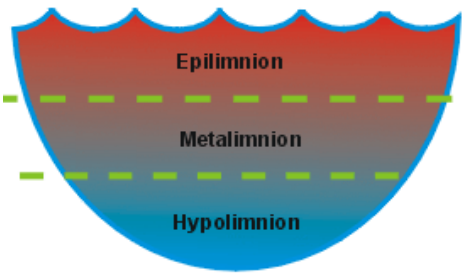


Cleveland & McGill, 1984 (*Journal of the American Statistical Association*)

15

Common Types of Graphs

- Lake Lacawac (Pennsylvania)
- Water temperature distribution



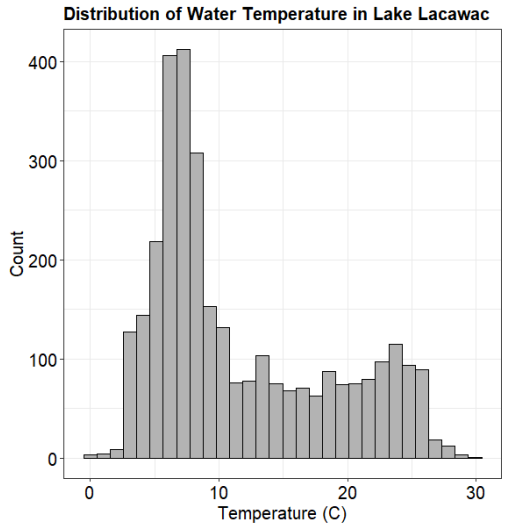
16

Common Types of Graphs

One variable:

- Temperature of Lake Lacawac

- Histogram



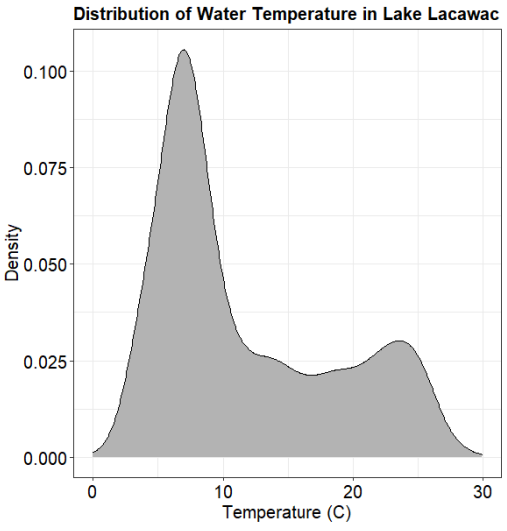
17

Common Types of Graphs

One variable:

- Temperature of Lake Lacawac

- Density plot



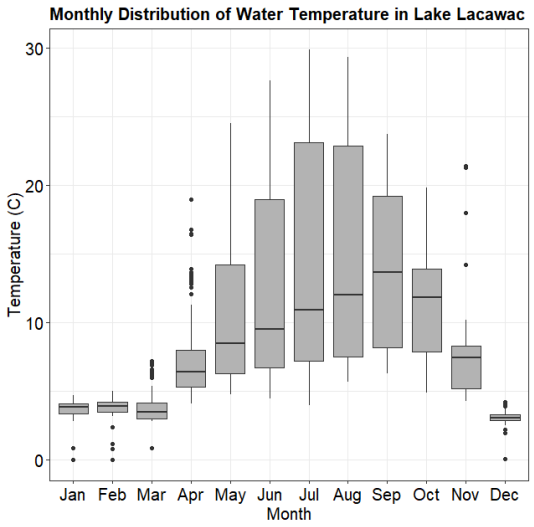
18

Common Types of Graphs

Two variables:

- Temperature of Lake Lacawac **for each month of the year**

- Box plot



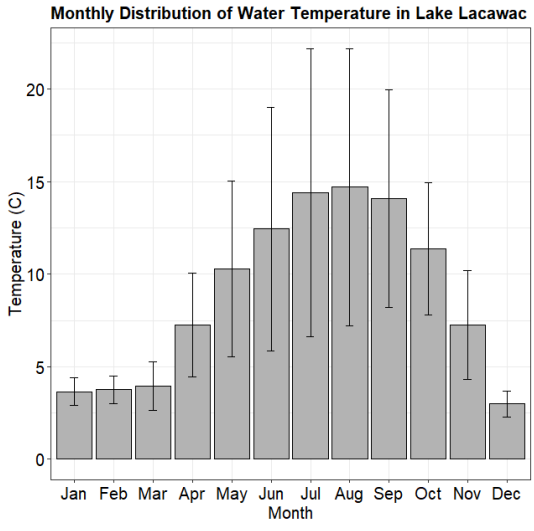
19

Common Types of Graphs

Two variables:

- Temperature of Lake Lacawac **for each month of the year**

- Bar plot



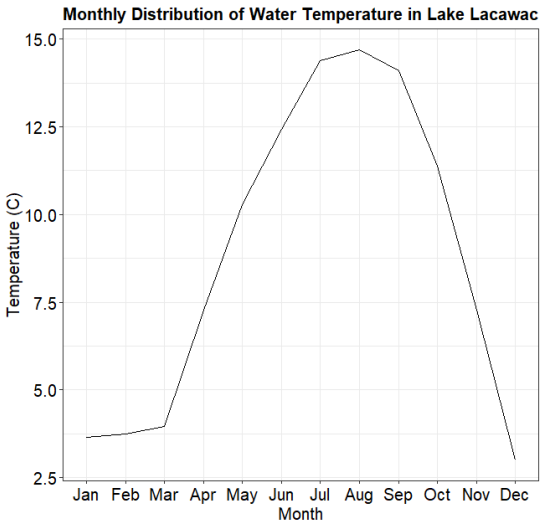
20

Common Types of Graphs

Two variables:

- Temperature of Lake Lacawac **for each month of the year**

- Line plot



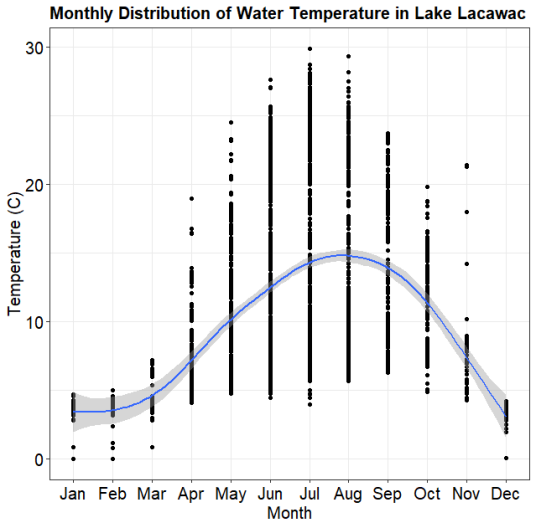
21

Common Types of Graphs

Two variables:

- Temperature of Lake Lacawac **for each month of the year**

- X-Y scatterplot



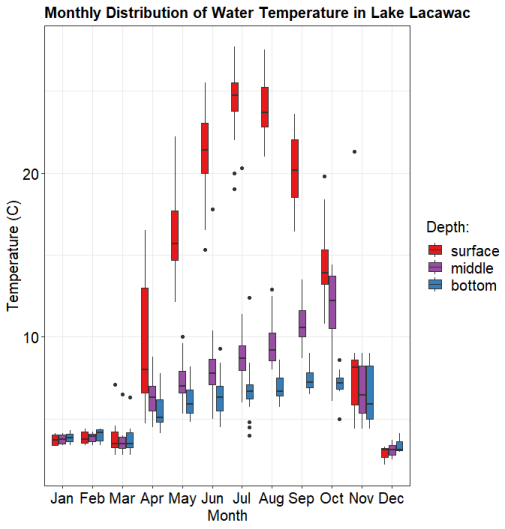
22

Common Types of Graphs

Three+ variables:

- Temperature of Lake Lacawac **<u>at multiple depths</u>** for each month of the year

- Box plot



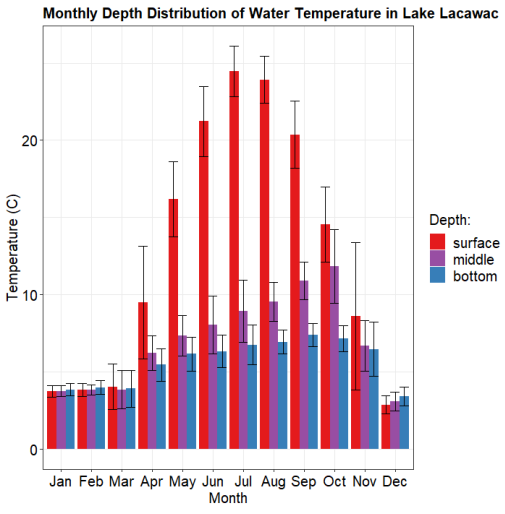
23

Common Types of Graphs

Three+ variables:

- Temperature of Lake Lacawac **<u>at multiple depths</u>** for each month of the year

- Bar plot



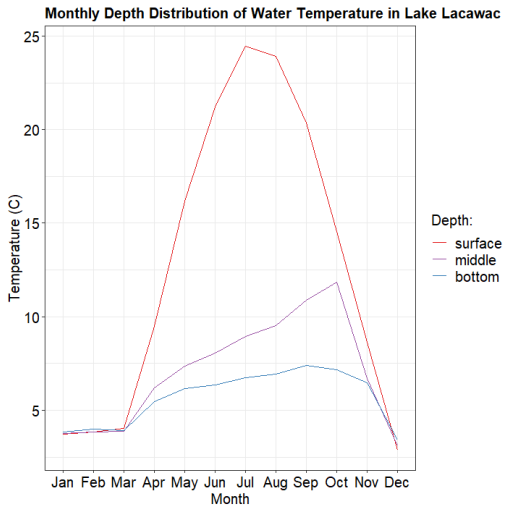
24

Common Types of Graphs

Three+ variables:

- Temperature of Lake Lacawac **<u>at multiple depths</u>** for each month of the year

- Line plot



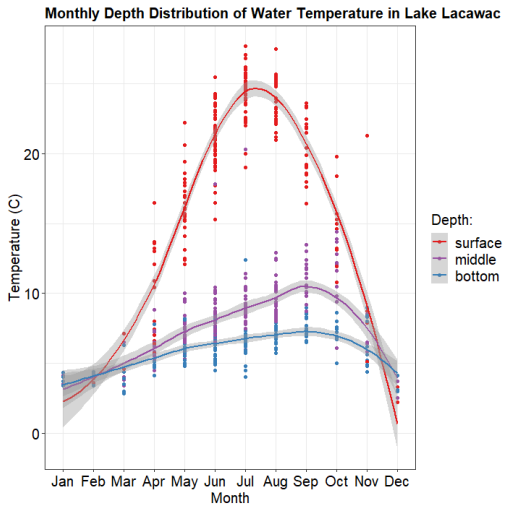
25

Common Types of Graphs

Three+ variables:

- Temperature of Lake Lacawac **<u>at multiple depths</u>** for each month of the year

- X-Y scatterplot



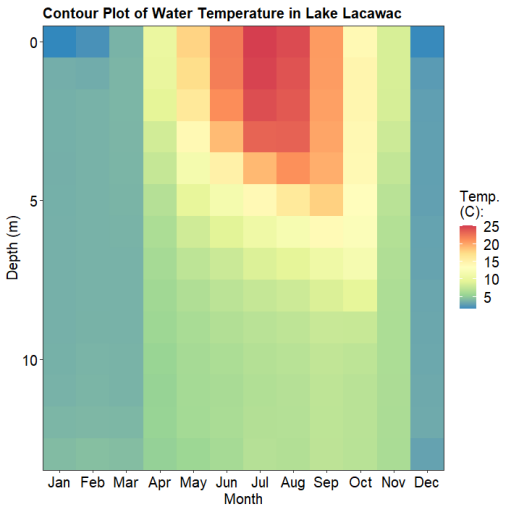
26

Common Types of Graphs

Three+ variables:

- Temperature of Lake Lacawac **<u>at multiple depths</u>** for each month of the year

- Contour plot



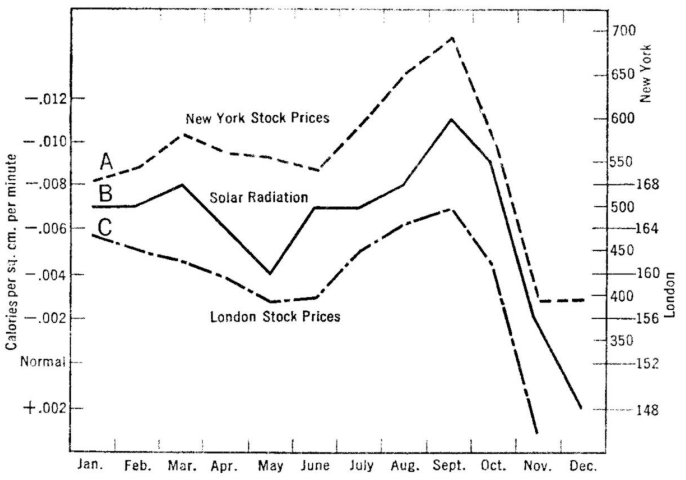
27

Keys to a Good Graph

- Graphs are only as useful as the data they display
- Don't accidentally or purposefully distort or mislead
- If the data is simple, ***keep the graph simple***
- If the data is complex, ***make the graph simple to understand***

28

Examples of Graphs



OVERLY COMPLEX GRAPH!

- Three axes with 3 different variables
- How is solar radiation directly related to stock prices?

from Ross Ihaka, University of Auckland, New Zealand

29

Examples of Graphs

UNNECESSARY GRAPH!

- Interpret 3 total components (2 axes and 1 color scheme) for only 2 pieces of data

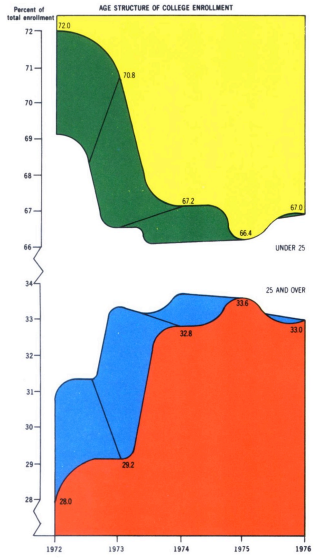
IMPROVEMENT:

- Just report "60% male and 40% female"

from Ross Ihaka, University of Auckland, New Zealand

30

Examples of Graphs



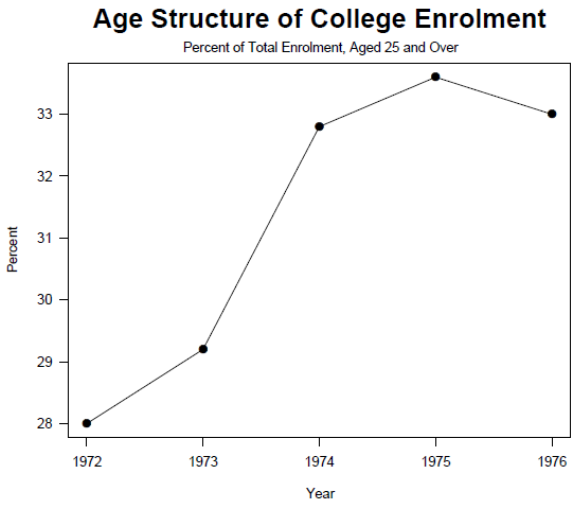
OVERLY COMPLEX GRAPH!

- Extra dimension is unnecessary – beware of 3-D graphs in Excel!
- Four colors misleads how many pieces of data are being presented
- Unnecessary splitting of y-axis

from Ross Ihaka, University of Auckland, New Zealand

31

Examples of Graphs



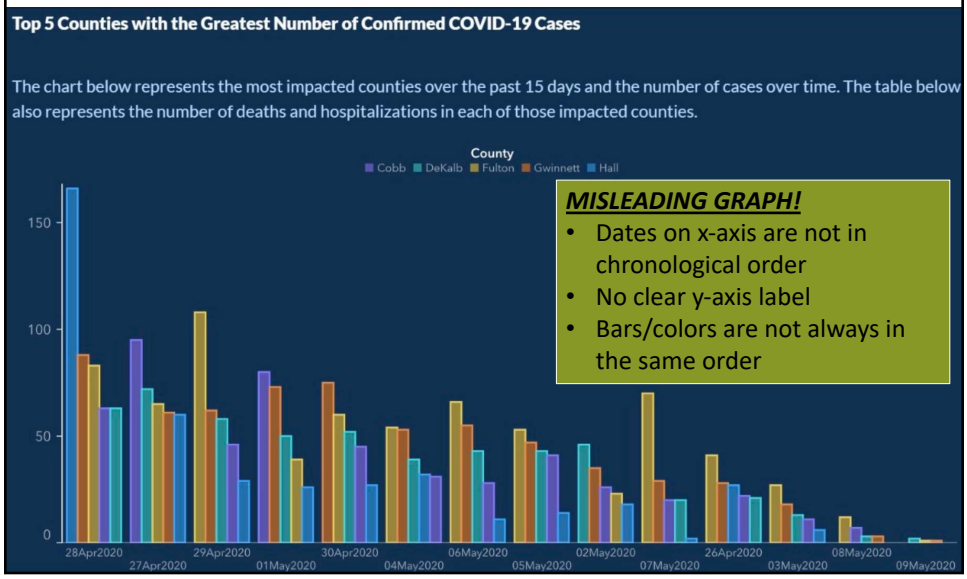
IMPROVEMENTS:

- Remove 3-D
- Focus on simple presentation of 5 key data points
- Clear axes

from Ross Ihaka, University of Auckland, New Zealand

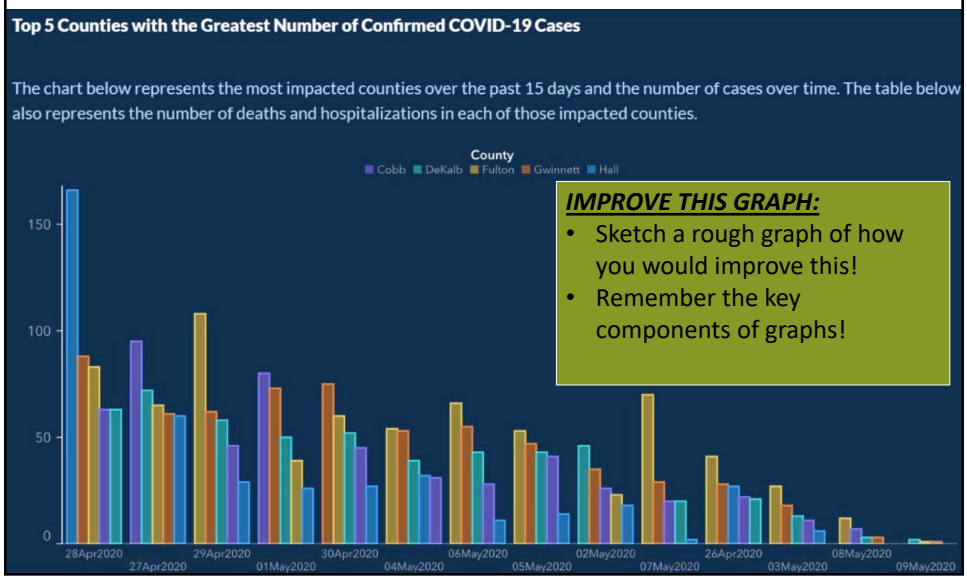
32

Examples of Graphs



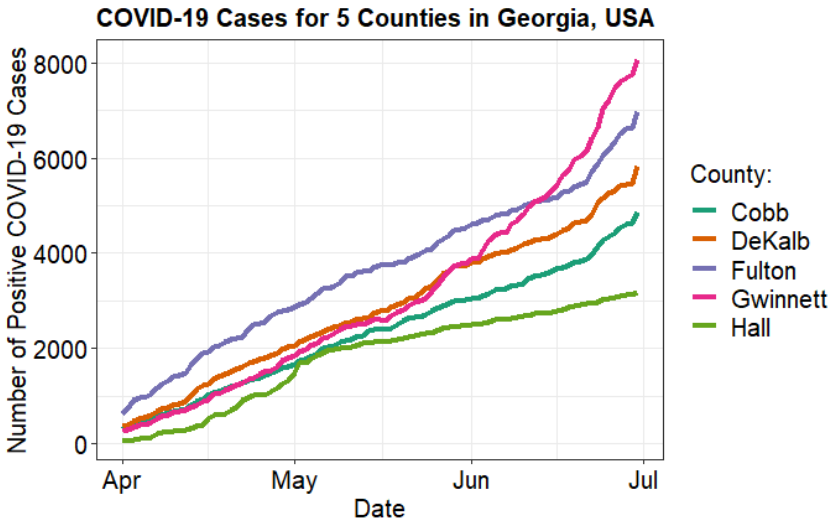
33

Examples of Graphs



34

Examples of Graphs



data from https://github.com/nytimes/covid-19-data

35

Live Coding

open RStudio and code along!

36

Overview of ggplot2

37

Grammar of Graphics

- Originally developed by Leland Wilkinson, adapted for R by Hadley Wickham
- Descriptive language to talk about components and visual elements of a graph or plot
  - Data being plotted
  - Geometric objects (lines, points, etc.)
  - Aesthetic appearance
  - Statistical information and transformations
  - Position, scales, coordinates
  - Facets or panels

38

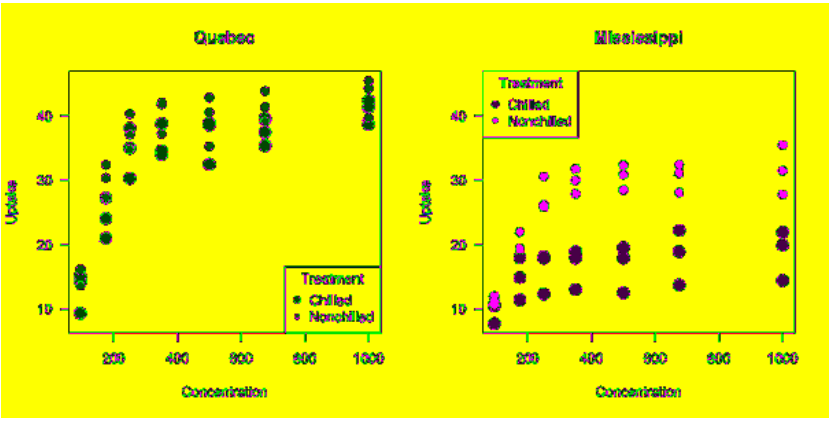
Grammar of Graphics

- Building a plot in ggplot2 is essentially constructing layers or pieces of a plot based on specific aspects of the data
- Each new layer is one new piece of code

39

Base R vs. ggplot2

CO_2 uptake in chilled vs. non-chilled treatments in 2 locations:



https://statsinthewild.com/2019/08/22/ggplot2-vs-base-r-graphics-an-example/

40

Base R vs. ggplot2

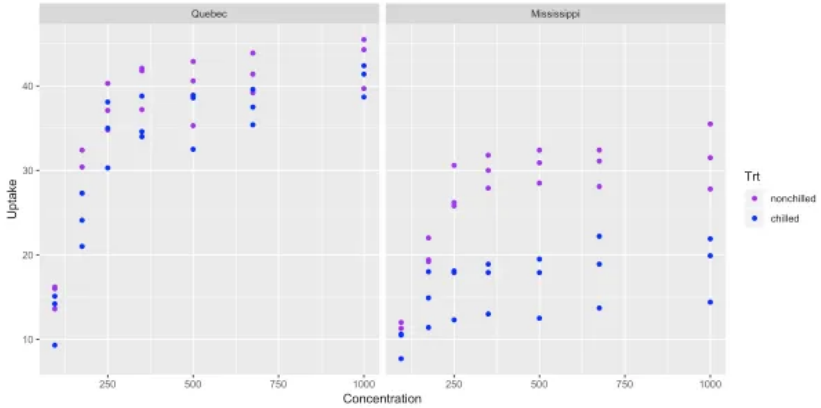
Base R code (30 lines):

```
summary(CO2)
str(CO2)
head(CO2, 20)
CO2.QC <- subset(CO2, Type == "Quebec" & Treatment == "chilled")
CO2.QN <- subset(CO2, Type == "Quebec" & Treatment == "nonchilled")
CO2.MC <- subset(CO2, Type == "Mississippi" & Treatment == "chilled")
CO2.MN <- subset(CO2, Type == "Mississippi" & Treatment == "nonchilled")
xrange <- range(CO2$conc)
yrange<-range(CO2$uptake)
par(mfrow = c(1, 2))
plot(CO2.QC$uptake ~ CO2.QC$conc, pch = 19, lty = 2, lwd = 3, cex = 1.5, las = 1,
ylim = yrange, xlim = xrange, col = "purple", xlab = "Concentration",
ylab = "Uptake")
par(new=T)
plot(CO2.QN$uptake ~ CO2.QN$conc, pch = 20, lty = 2, lwd = 3, cex = 1.5, las = 1,
ylim = yrange, xlim = xrange, col = "lightblue", main = "Quebec", xlab = "",
ylab = "")
legend("bottomright", title = "Treatment", c("Chilled", "Nonchilled"),
pch = c(19, 20), col = c("purple", "lightblue"))
plot(CO2.MC$uptake ~ CO2.MC$conc, pch = 19, lty = 2, lwd = 3, cex = 1.5, las = 1,
ylim = yrange, xlim = xrange, col = "orange", xlab = "Concentration",
ylab = "Uptake")
par(new=T)
plot(CO2.MN$uptake ~ CO2.MN$conc, pch = 20, lty = 2, lwd = 3, cex = 1.5,
ylim = yrange, las = 1,
xlim = xrange, col = "red", main = "Mississippi", xlab = "", ylab = "")
legend("topleft", title = "Treatment", c("Chilled", "Nonchilled"), pch = c(19,20),
col = c("orange", "red"))
par(mfrow = c(1, 1), las = 1)
```

41

Base R vs. ggplot2

CO_2 uptake in chilled vs. non-chilled treatments in 2 locations:



https://statsinthewild.com/2019/08/22/ggplot2-vs-base-r-graphics-an-example/

42

Base R vs. ggplot2

ggplot2 code (8 lines):

```
library(ggplot2)
ggplot(aes(x = conc, y = uptake, color = Treatment), data = CO2) +
  geom_point() +
  facet_grid( ~ Type) +
  xlab("Concentration") +
  ylab("Uptake") +
  labs(color = "Trt") +
  scale_color_manual(values = c("purple","blue"))
```

43

Base R vs. ggplot2

Plotting in base R can be very useful for…

1) Beginners learning R
2) Exploratory plots (don't have to be perfect)
3) Heat or contour maps with missing data

ggplot2 is often much quicker and more intuitive when creating…

1) Legends
2) Groupings
3) Faceting or panels
4) Easy customizing for presentations or publications

44

Computer Science in Modern Biology

Data Visualization in R

Instructor:
Rachel Pilla

TAs:
Mason Murphy

MIAMI

Mahbobeh Lesani

45

Review of Day 1

46

Workshop Logistics

Zoom
- Feel free to unmute yourself to ask questions at any point

Chat
- Post any questions or comments here
- The TAs will be checking the chat to help answer questions

Break-Out Rooms
- If you are really stuck, you can move to a break-out room with a TA to share your screen and get you up to speed
- Break time is a great opportunity for this if needed!

47

Packages

- Packages only need to be installed ONCE
- But, each time you re-open R (***today!***), you need to load the packages you want to use so R can make all the functions available:

```
library(package name)
```

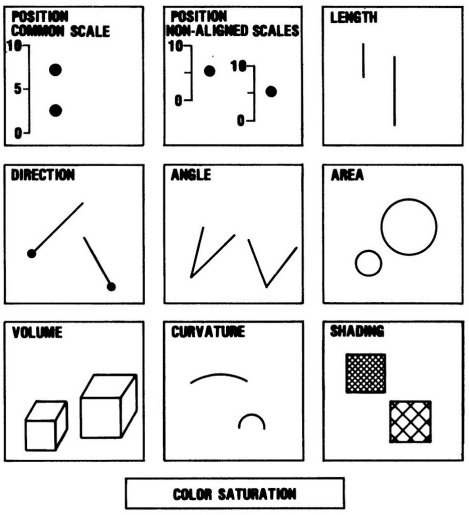
48

Components of a Graph

1) Axes
   - Horizontal x-axis and vertical y-axis
2) Labels
   - What are your axes measuring?
3) Title and/or caption
   - Description and information about the graph
4) Legend (sometimes)
   - If multiple lines, colors, shapes, etc., what does each represent?
5) DATA
   - The actual information!

49

Understanding a Graph



- Easy to perceive and understand

- May require more work to understand
- Can be very valuable when used appropriately in graphs

Cleveland & McGill, 1984 (*Journal of the American Statistical Association*)

50

Remember: Keys to a Good Graph

- Graphs are only as useful as the data they display
- Don't accidentally or purposefully distort or mislead
- If the data is simple, ***keep the graph simple***
- If the data is complex, ***make the graph simple to understand***

51

Introduction to ggplot2

52

Grammar of Graphics

- Originally developed by Leland Wilkinson, adapted for R by Hadley Wickham
- Descriptive language to talk about components and visual elements of a graph or plot
  - Data being plotted
  - Geometric objects (lines, points, etc.)
  - Aesthetic appearance
  - Statistical information and transformations
  - Position, scales, coordinates
  - Facets or panels

53

Grammar of Graphics

- Building a plot in ggplot2 is essentially constructing layers or pieces of a plot based on specific aspects of the data

  Layer 1: Plot points of water temperature (y-axis) over time (x-axis)
  Layer 2: Scale axes to preference
  Layer 3: Add labels for axes and plot title
  Layer 4: Make adjustments for font size, etc.

- Each new layer is one new piece of code

54

Grammar of Graphics

```
ggplot() +
```

Base Layer:
Initiate ggplot2 (always start a new plot with this code)

55

Grammar of Graphics

```
ggplot() +
```

"+" indicates another layer will follow

56

Grammar of Graphics

```
ggplot() +
  geom_point( … ) +
```

Layer 1:
Plot points of water temperature (y-axis) over time (x-axis)

57

Grammar of Graphics

```
ggplot() +
  geom_point( … ) +
  scale_x_continuous( … ) +
```

Layer 2:
Scale axes to preference

58

Grammar of Graphics

```
ggplot() +
  geom_point( … ) +
  scale_x_continuous( … ) +
  labs( … ) +
```

Layer 3:
Add labels for axes and plot title

59

Grammar of Graphics

```
ggplot() +
  geom_point( … ) +
  scale_x_continuous( … ) +
  labs( … ) +
  theme( … )
```

Layer 4:
Make adjustments for font size, etc.

60

Geometries

- Describe the different types or 'shapes" of graphs that will be produced, and are very self-explanatory
- Examples:
  - geom_point
  - geom_line
  - geom_histogram
  - geom_smooth
  - geom_col
  - geom_bar

61

Aesthetic

- Indicates which variables will be plotted and on which axis or scale
- Inside the geometry function, always use the following coding syntax:

```
geom_point( mapping = aes( … ), … )
```

- Use column names to map variables to x, y, color, fill, etc.

62

Options & Customizations

- Additional layers to make the graph look how you want it to

- Examples:
  - Labels
  - Color or fill scales
  - Axis scales
  - Themes
  - Annotations
  - Facets or panels
  - Many other options!

63

Live Coding

open RStudio and code along!

64

Q&A Session

65

Additional Tips & Useful Resources

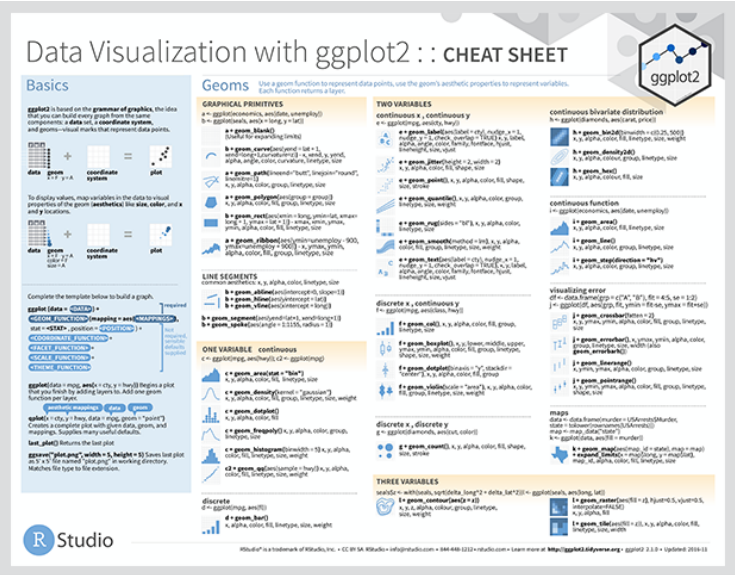
66

RStudio Cheat Sheets

- Short, handy guides to help with many packages and programming needs in RStudio
- Available free for download in multiple language options
- https://rstudio.com/resources/cheatsheets/

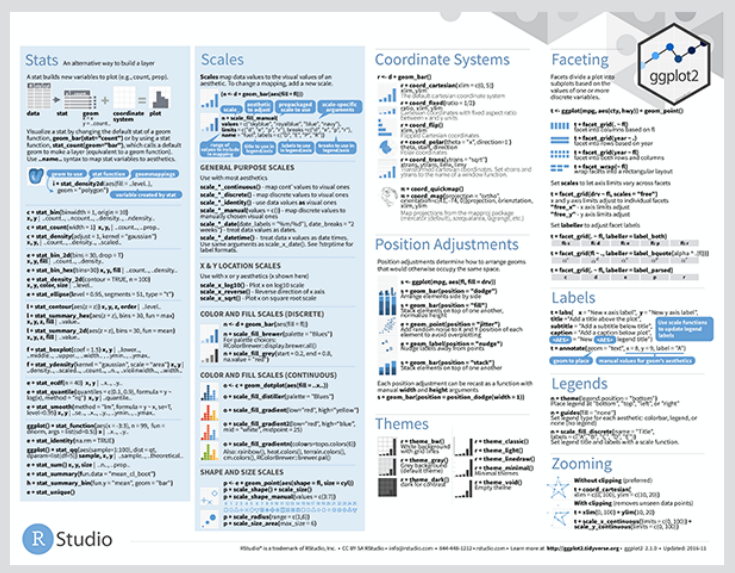
67

ggplot2 Cheat Sheet



68

ggplot2 Cheat Sheet



69

Tidyverse

- Packages to help organize data and streamline R coding, specifically for data science
- Functions in the packages share similar syntax and coding structure
- “dplyr”, “tidyr”, and “ggplot2” are part of tidyverse
- https://www.tidyverse.org/

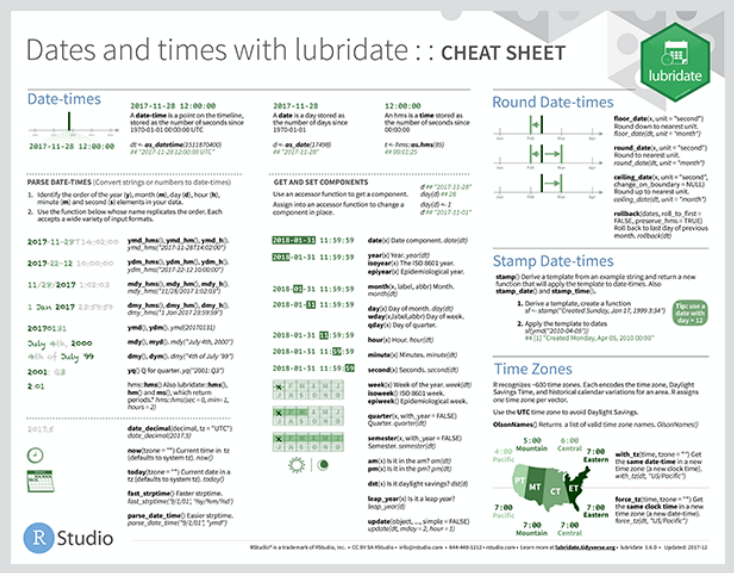
70

Lubridate

- Package for more effectively working with dates between Excel and R
- Can recognize date/times with minimal formatting issues
- Also allows for date/time differencing, among other useful functions
- Included in tidyverse
- https://lubridate.tidyverse.org/

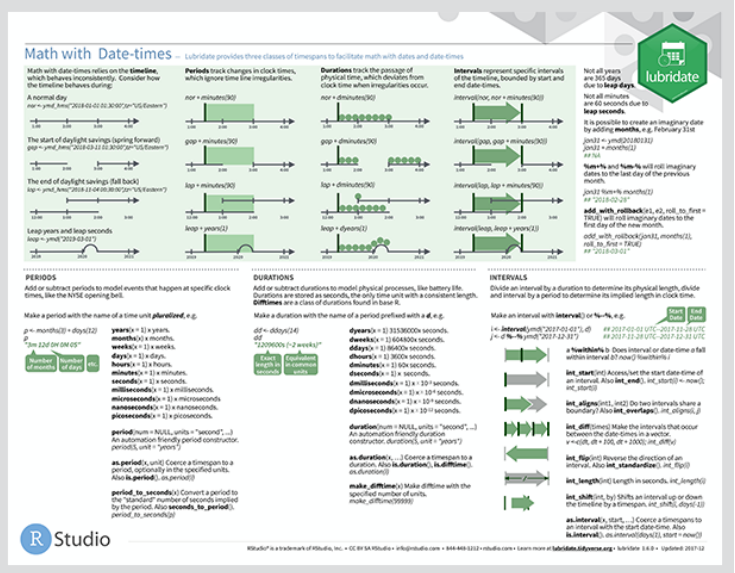
71

Lubridate



72

Lubridate



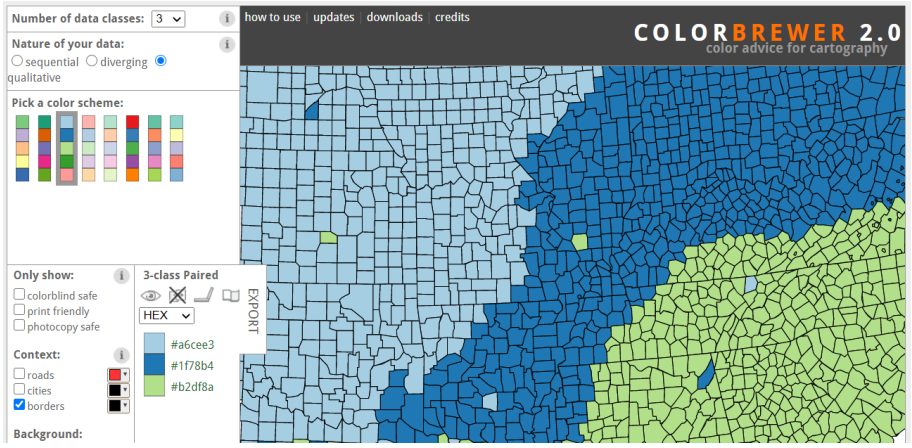
73

ColorBrewer

- Pre-made color schemes and palettes
- Effective graphic color selections for maximum perception
  - Qualitative, sequential, and diverging
- Includes options for photo-copy friendly, color-blind friendly, etc.
- http://colorbrewer2.org/

74

ColorBrewer



75

Swirl

- Package for self-guided learning of R in the R environment
- Learn at your own pace and select topics most useful to you
  - Introductory material
  - Regressions
  - Data wrangling
  - Statistical inference
  - Advanced programming
- https://swirlstats.com/

76

Thank you for joining!

to share feedback or ask questions,
please feel free to contact me:

pillarm@miamioh.edu
@rmpilla

77

Notes