11: Crafting Reports

Stephanie Kinser

Spring 2022

LESSON OBJECTIVES

1. Describe the purpose of using R Markdown as a communication and workflow tool
2. Incorporate Markdown syntax into documents
3. Communicate the process and findings of an analysis session in the style of a report

USE OF R STUDIO & R MARKDOWN SO FAR...

1. Write code
2. Document that code
3. Generate PDFs of code and its outputs
4. Integrate with Git/GitHub for version control

BASIC R MARKDOWN DOCUMENT STRUCTURE

1. **YAML Header** surrounded by — on top and bottom
    - YAML templates include options for html, pdf, word, markdown, and interactive
    - More information on formatting the YAML header can be found in the cheat sheet
2. **R Code Chunks** surrounded by "`on top and bottom  + Create using`Cmd/Ctrl+Alt+I'
    - Can be named {r name} to facilitate navigation and autoreferencing
    - Chunk options allow for flexibility when the code runs and when the document is knitted
3. **Text** with formatting options for readability in knitted document

RESOURCES

Handy cheat sheets for R markdown can be found: here, and here.

There's also a quick reference available via the `Help→Markdown Quick Reference` menu.

Lastly, this website give a great & thorough overview.

THE KNITTING PROCESS

- The knitting sequence 

- Knitting commands in code chunks:
- `include = FALSE` - code is run, but neither code nor results appear in knitted file
- `echo = FALSE` - code not included in knitted file, but results are
- `eval = FALSE` - code is not run in the knitted file
- `message = FALSE` - messages do not appear in knitted file
- `warning = FALSE` - warnings do not appear...
- `fig.cap = "..."` - adds a caption to graphical results

WHAT ELSE CAN R MARKDOWN DO?

See: https://rmarkdown.rstudio.com and class recording. * Languages other than R... * Various outputs...

WHY R MARKDOWN?

<Fill in our discussion below with bullet points. Use italics and bold for emphasis (hint: use the cheat sheets or `Help` →`Markdown Quick Reference` to figure out how to make bold and italic text).>

R Markdown excels in the following:

- *Team collaboration*
  - Allows for **version control** via Git
  - Can easily **run chunks** of collaborators' code
- *Easily edit reports*
  - Change code and rerun reports
  - Integrate text and code
- *Beginner-friendly*
  - See code, output, files, plots all in one place
  - **Packages** make it easy to start coding

TEXT EDITING CHALLENGE

Create a table below that details the example datasets we have been using in class. The first column should contain the names of the datasets and the second column should include some relevant information about the datasets. (Hint: use the cheat sheets to figure out how to make a table in Rmd)

Data Set Name	Description
EPAAir	EPA air quality data
OzoneTimeSeries	Ozone concentration time series
NTL-LTER_Lake_Nutrients	Lake nutrient data
ECOTOX	Ecotoxicology data
NIWO_Litter	NIWO litter data

R CHUNK EDITING CHALLENGE

Installing packages

Create an R chunk below that installs the package `knitr`. Instead of commenting out the code, customize the chunk options such that the code is not evaluated (i.e., not run).

```
install.packages('knitr')
```

Setup

Create an R chunk below called "setup" that checks your working directory, loads the packages `tidyverse`, `lubridate`, and `knitr`, and sets a ggplot theme. Remember that you need to disable R throwing a message, which contains a check mark that cannot be knitted.

```
getwd()
```

```
## [1] "C:/Users/pogo/Documents/ENV872/Environmental_Data_Analytics_2022/Assignments"
```

```
library(lubridate)
library(knitr)
library(tidyverse)

mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

Load the NTL-LTER_Lake_Nutrients_Raw dataset, display the head of the dataset, and set the date column to a date format.

Customize the chunk options such that the code is run but is not displayed in the final document.

Data Exploration, Wrangling, and Visualization

Create an R chunk below to create a processed dataset do the following operations:

- Include all columns except lakeid, depth_id, and comments
- Include only surface samples (depth = 0 m)
- Drop rows with missing data

```
##          lakename year4 daynum sampledate depth tn_ug tp_ug   nh34   no23 po4
## 5       Paul Lake  1991    147 1991-05-27     0   483     8 18.981 10.299   8
## 9       Paul Lake  1991    154 1991-06-03     0   512    17 20.736  5.934   4
## 10     Peter Lake  1991    155 1991-06-04     0   573     5 21.533  3.818   4
## 12 West Long Lake  1991    157 1991-06-06     0   490     4 26.585  2.465   4
## 13      Paul Lake  1991    161 1991-06-10     0   421    11 20.303 11.381   6
## 14     Peter Lake  1991    162 1991-06-11     0   296     9 14.579  6.827   4
```

Table 2: Nitrogen Summary Statistics

Lake Name	Mean	Median	Standard Deviation
Central Long Lake	690.0	751	209.0
Crampton Lake	362.7	358	12.1
East Long Lake	810.8	763	335.0
Hummingbird Lake	1037.0	1070	204.0
Paul Lake	368.8	383	106.0
Peter Lake	561.9	511	306.0
Tuesday Lake	423.6	433	78.8
West Long Lake	762.6	693	403.0

Table 3: Phosphorus Summary Statistics

Lake name	Mean	Median	Standard Deviation
Central Long Lake	21.71	21.70	7.08
Crampton Lake	11.16	12.10	4.95
East Long Lake	29.29	24.70	17.40
Hummingbird Lake	36.22	35.00	4.15
Paul Lake	10.46	9.51	4.81
Peter Lake	18.39	16.80	11.00
Tuesday Lake	11.72	12.10	3.04
West Long Lake	19.83	18.10	10.50

Create a second R chunk to create a summary dataset with the mean, minimum, maximum, and standard deviation of total nitrogen concentrations for each lake. Create a second summary dataset that is identical except that it evaluates total phosphorus. Customize the chunk options such that the code is run but not displayed in the final document.

Create a third R chunk that uses the function `kable` in the knitr package to display two tables: one for the summary dataframe for total N and one for the summary dataframe of total P. Use the `caption = " "` code within that function to title your tables. Customize the chunk options such that the final table is displayed but not the code used to generate the table.

Create a fourth and fifth R chunk that generates two plots (one in each chunk): one for total N over time with different colors for each lake, and one with the same setup but for total P. Decide which geom option will be appropriate for your purpose, and select a color palette that is visually pleasing and accessible. Customize the chunk options such that the final figures are displayed but not the code used to generate the figures. In addition, customize the chunk options such that the figures are aligned on the left side of the page. Lastly, add a fig.cap chunk option to add a caption (title) to your plot that will display underneath the figure.

Communicating results

Write a paragraph describing your findings from the R coding challenge above. This should be geared toward an educated audience but one that is not necessarily familiar with the dataset. Then insert a horizontal rule below the paragraph. Below the horizontal rule, write another paragraph describing the next steps you might take in analyzing this dataset. What questions might you be able to answer, and what analyses would you conduct to answer those questions?

> Findings: The nitrogen and phosphorous concentrations in all lakes varied within each year and across years. Through the 1990s, there is a general upward trend in concentrations for both chemicals. Accounting for all years, Hummingbird Lake has the highest mean and median concentrations of both nitrogen and phosphorous. Conversely, Crampton Lake has the lowest

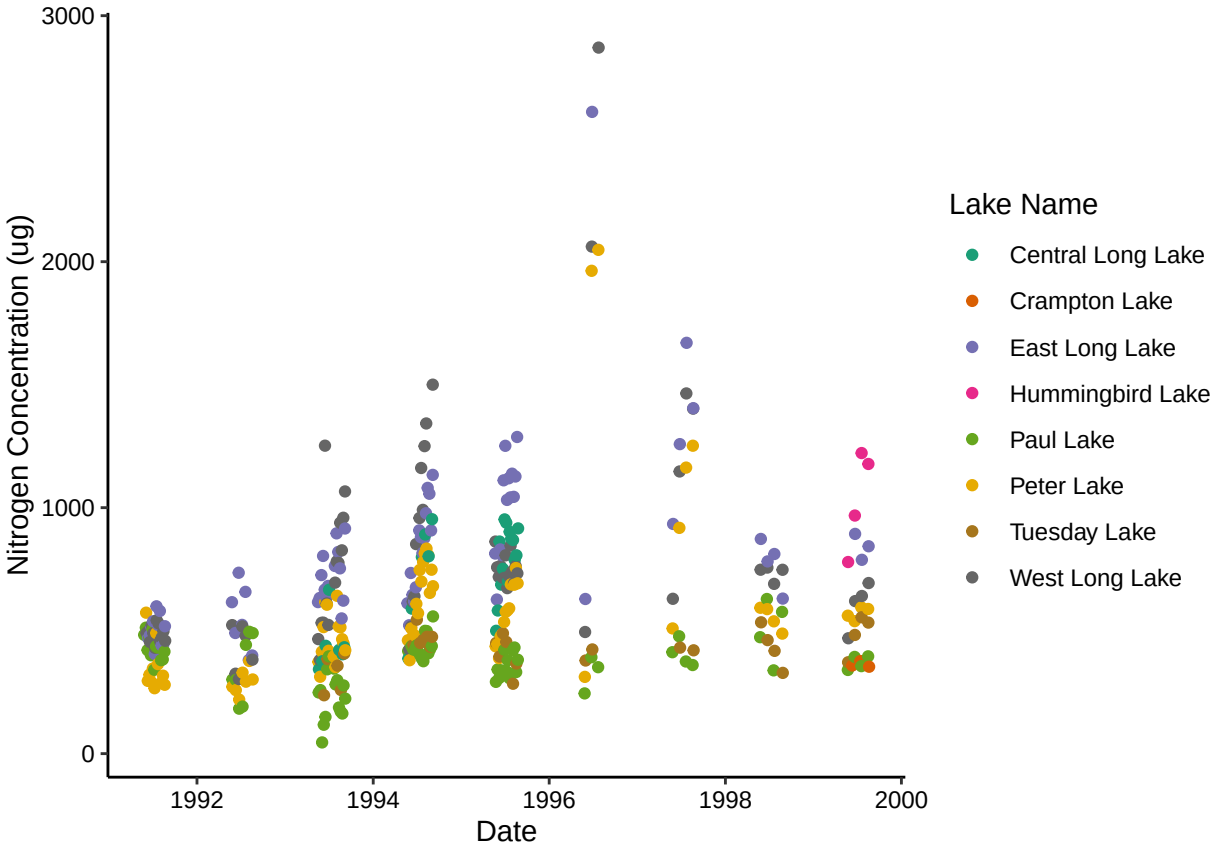

Figure 1: Nitrogen Concentrations Time Series

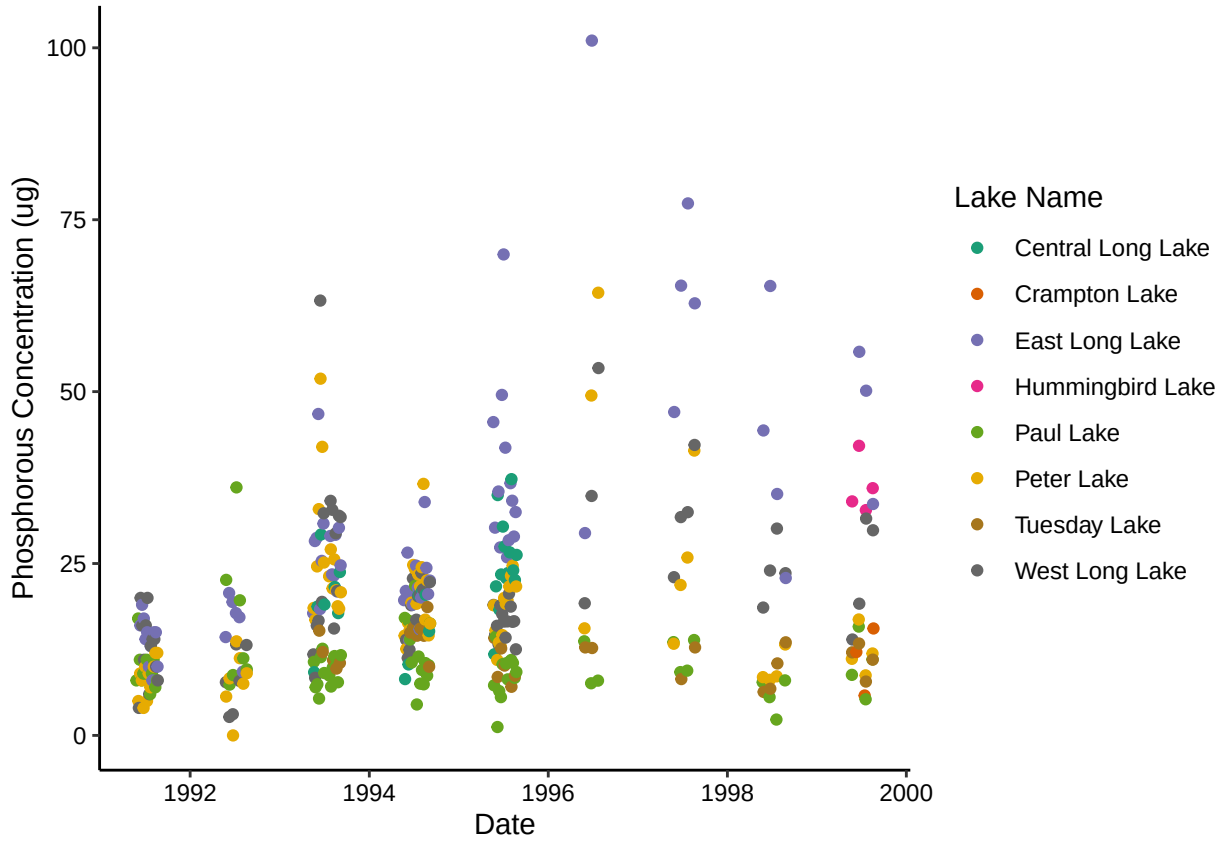

Figure 2: Phosphorous Concentrations Time Series

mean nitrogen concentration and Paul Lake has the lowest mean phosphorous concentration across all years.

Next Steps: To more fully analyze the chemical concentrations of the lakes, I would consider comparing concentrations for each month across years to better assess seasonal changes in the chemical composition of the lakes. For this assessment, I would run a time series analysis. I would also consider running linear regression models to analyze whether there are significant relationships between the different chemicals found in the lakes and compare these results across the lakes.

KNIT YOUR PDF

When you have completed the above steps, try knitting your PDF to see if all of the formatting options you specified turned out as planned. This may take some troubleshooting.

OTHER R MARKDOWN CUSTOMIZATION OPTIONS

We have covered the basics in class today, but R Markdown offers many customization options. A word of caution: customizing templates will often require more interaction with LaTeX and installations on your computer, so be ready to troubleshoot issues.

Customization options for pdf output include:

- Table of contents
- Number sections
- Control default size of figures
- Citations
- Template (more info here)

pdf_document:
toc: true
number_sections: true
fig_height: 3
fig_width: 4
citation_package: natbib
template: