11: Crafting Reports

Environmental Data Analytics | John Fay & Luana Lima | Developed by Kateri Salk

Spring 2021

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

Figure 1: knitting

- `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).>

- Create clean documents with text accompanied by code and outputs, including figures, to **share with collaborators**
- Include chunks from **other coding languages** (*if you have the kernels installed*)
- Create **multiple types of documents** including PDFs, HTML, and presentations.
- **Customize settings** for which code chunks are run and whether outputs and messages are displayed
- **Navigate easily** by naming code chunks and defining sections
- Using text editing, **track changes** easily through *version control*

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)

```
Title <- c("EPA Air NC (e.g. EPAair_O3_NC2018_raw.csv)",
           "NEON NIWO Litter (e.g. NEON_NIWO_Litter_trapdata_raw.csv)",
           "NTL LTER Lake (e.g. NTL-LTER_Lake_Nurtients_Raw.csv)",
           "NWIS Flow Data (e.g. NWIS_SiteFlowData_NE_RAW.csv)")
Description <- c("Air quality data (O3 and PM2.5) for NC, USA",
                 "Leaf litter mass and functional group from Niwot Ridge sites",
                 "Lake depth, temperature, and nutrient data from North Temperate Lakes Long-Term Ecolo
```

```
                "Stream gage data from National Water Information System")
dataset_info <- data.frame(Title, Description)
knitr::kable(dataset_info, caption = "Table of dataset info")
```

Table 1: Table of dataset info

Title	Description
EPA Air NC (e.g. EPAair_O3_NC2018_raw.csv)	Air quality data (O3 and PM2.5) for NC, USA
NEON NIWO Litter (e.g. NEON_NIWO_Litter_trapdata_raw.csv)	Leaf litter mass and functional group from Niwot Ridge sites
NTL LTER Lake (e.g. NTL-LTER_Lake_Nurtients_Raw.csv)	Lake depth, temperature, and nutrient data from North Temperate Lakes Long-Term Ecological Research project
NWIS Flow Data (e.g. NWIS_SiteFlowData_NE_RAW.csv)	Stream gage data from National Water Information System

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/Anne/Documents/ENV872/Environmental_Data_Analytics_2021/Lessons"
```

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

mytheme <- theme_bw() + theme(
  text = element_text(color = "black", size = 12),
  axis.text = element_text(color = "black", size = 12),
  legend.position = "top",
  axis.line = element_line(color = "dark gray", size = 1.25))
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

```
lakedata_processed <- lakedata %>%
    select(!lakeid, !depth_id, !comments) %>%
    filter(depth == "0") %>%
    drop_na()
```

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.

Table 2: Summary of total nitrogen (ug) by lake

Lake Name	Mean total N	Minimum total N	Maximum total N	Standard deviation total N
Central Long Lake	690.0469	343.020	953.063	209.09341
Crampton Lake	362.6813	353.380	376.304	12.05748
East Long Lake	810.7834	380.620	2608.956	335.41457
Hummingbird Lake	1036.6695	779.053	1221.960	204.36889
Paul Lake	368.7564	45.670	628.625	106.34741
Peter Lake	561.8752	219.720	2048.151	305.64909
Tuesday Lake	423.5605	237.363	554.418	78.84522
West Long Lake	762.6017	303.170	2870.302	402.95992

Table 3: Summary of total phosphorus (ug) by lake

Lake Name	Mean total P	Minimum total P	Maximum total P	Standard deviation total P
Central Long Lake	21.70981	8.190	37.270	7.076388
Crampton Lake	11.16033	5.803	15.555	4.946759
East Long Lake	29.28984	8.000	101.050	17.375710
Hummingbird Lake	36.21925	32.765	42.119	4.146717
Paul Lake	10.45606	1.222	36.070	4.805142
Peter Lake	18.39153	0.000	64.383	10.976205
Tuesday Lake	11.71853	6.325	18.663	3.044289
West Long Lake	19.82981	2.690	63.243	10.541276

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.

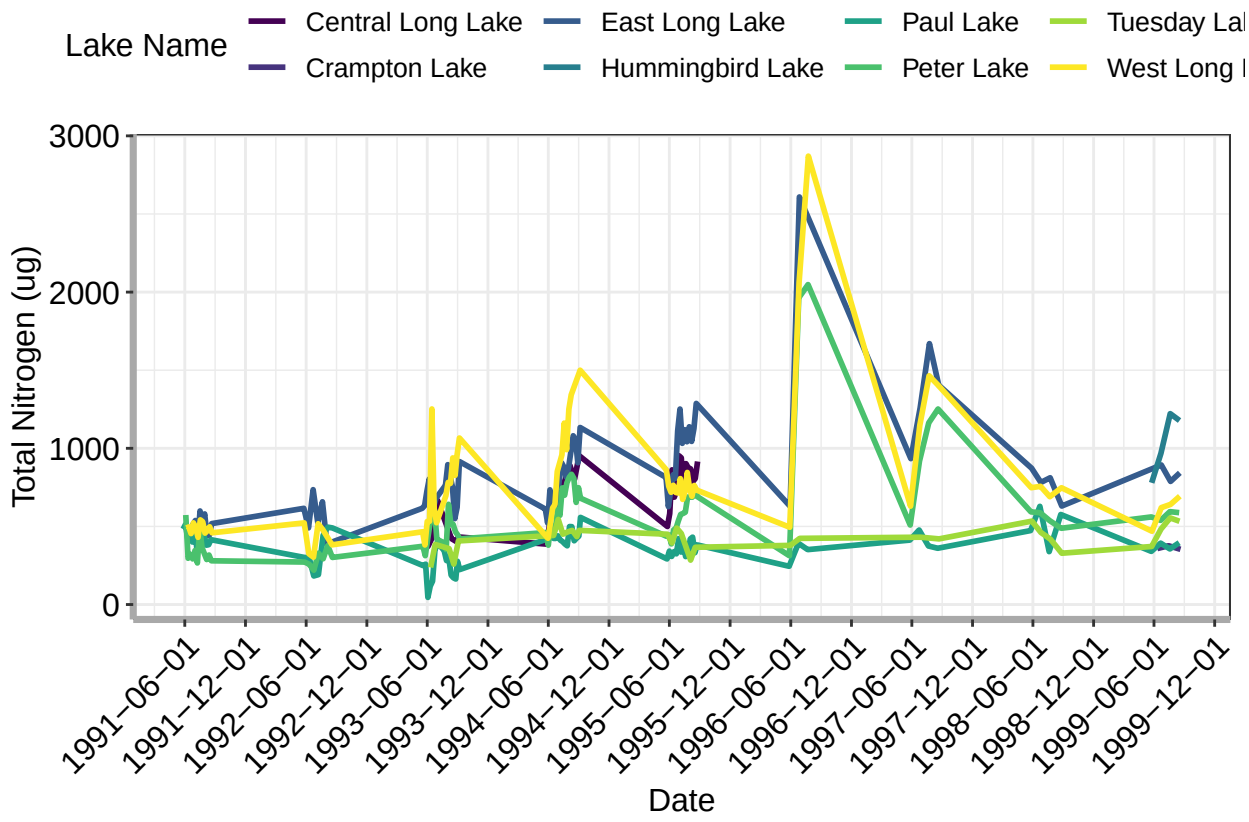

Figure 2: Total nitrogen by lake for NTL LTER lakes from 1991-1999

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?

On average, Hummingbird Lake has the highest nutrient levels (both nitrogen and phosphorus; Tables 2 and 3 respectively) of any lake in the Northern Temperate Lakes Long-Term Ecological Research program, with an average of 1036.67 ug of nitrogen and 36.22 ug of phosphorus recorded at the surface. Meanwhile, Crampton Lake has the lowest average level of total nitrogen (362.68 ug) and Paul Lake has the lowest average level of phosphorus (10.46 ug). Nitrogen levels were most variable in West Long Lake (range = 2567.13, sd = 402.96) and phosphorus levels were most variable in East Long Lake (range = 93.05, sd = 17.38). The levels of total nitrogen (Figure 1) and total phosphorus (Figure 2) in the lakes of the NTL-LTER fluctuate over time. Each year, nutrients are relatively low at the beginning of summer (late May to early June) and higher at the end of summer (late July to early August). In Peter Lake and East, Central, and West Long Lake, the level of both nitrogen and phosphorus appears to increase over time. In general, it

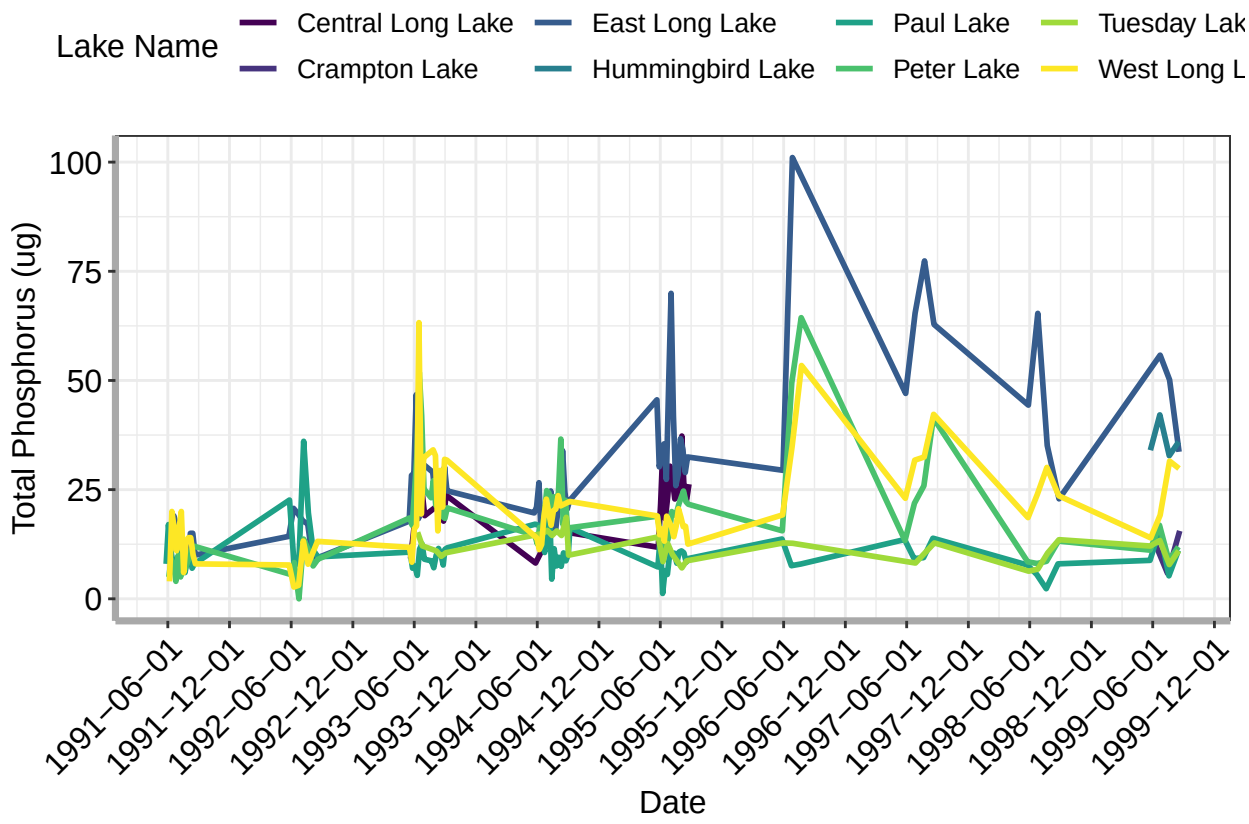

Figure 3: Total phosphorus by lake for NTL LTER lakes from 1991-1999

appears that the level of nitrogen and phosphorus are correlated; peaks in nitrogen level usually correspond to peaks in phosphorus level.

To better understand patterns in nutrient levels in the LTER lakes, I would begin by investigating whether the nitrogen and phosphorus levels are significantly different in different lakes. I could answer this question with a one-way analysis of variance (ANOVA) for total nitrogen and another for total phosphorus to compare nutrient levels among lakes. If the ANOVA returns a signifcant result, I will follow up with Tukey's HSD pairwise comparisons to see which lakes have significantly different nutrient levels from the others. Next, I would use a time series analysis to investigate whether there is a seasonal trend in nutrient levels and a trend across years, as the two figures suggest. Two challenges to conducting a time series analysis would be 1. the period of data collection, as data on lake nutrient levels is only collected during the summer; and 2. the temporal coverage is not equal for all lakes - for example, Hummingbird Lake only has data recorded in 1999. If the time series analysis is not a good fit for the data, I could also attempt to answer this question using a linear regression with month and/or year as explanatory variables and nutrient level as the response variable.

Duke Community Standard affirmation: I have adhered to the Duke Community Standard in completing this assignment. -Anne Harshbarger

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: