11: Crafting Reports

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

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

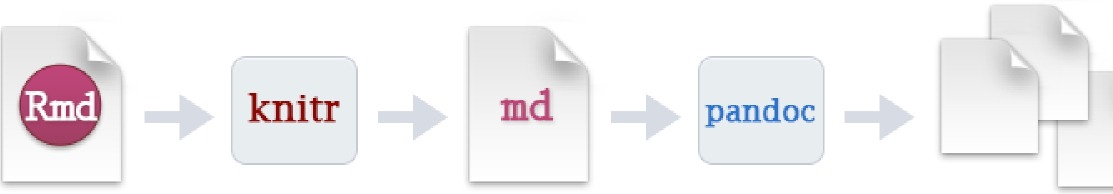
Handy cheat sheets for R markdown can be found: here, and here.

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… + Can do presentations!!

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).>

- Can take notes more clearly, add sections, add equations
- **Much** more accessible
  - *As compared to writing plain code*
- Can present code and results, without requiring others to run your code
- Can create full document/report within R without having to copy/paste, re-run, etc.
- RStudio seamless connection with github allows for great version control with git connection
- Can use many languages

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)

Dataset Name	Dataset Details
EPA Air Quality Datasets	These datasets include PM2.5 and Ozone data from NC in 2017/18.
NEON Niwot Ridge litter dataset	This is litter and woody debris data from the Niwot Ridge Long-Term Ecological Research.

Dataset Name	Dataset Details
NTL-LTER Lake Dataset	This data is from several lakes in the North Temperate Lakes District in Wisconsin, USA.

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).

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.

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*

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 Nitrogen Concentrations in NTR LTER Lakes

Mean	Maximum	Minimum	Standard Deviation
610.9982	2870.302	45.67	333.8124

Table 3: Summary of Phosphorous Concentrations in NTR LTER Lakes

Mean	Maximum	Minimum	Standard Deviation
3.28626	62.535	0	4.684112

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.

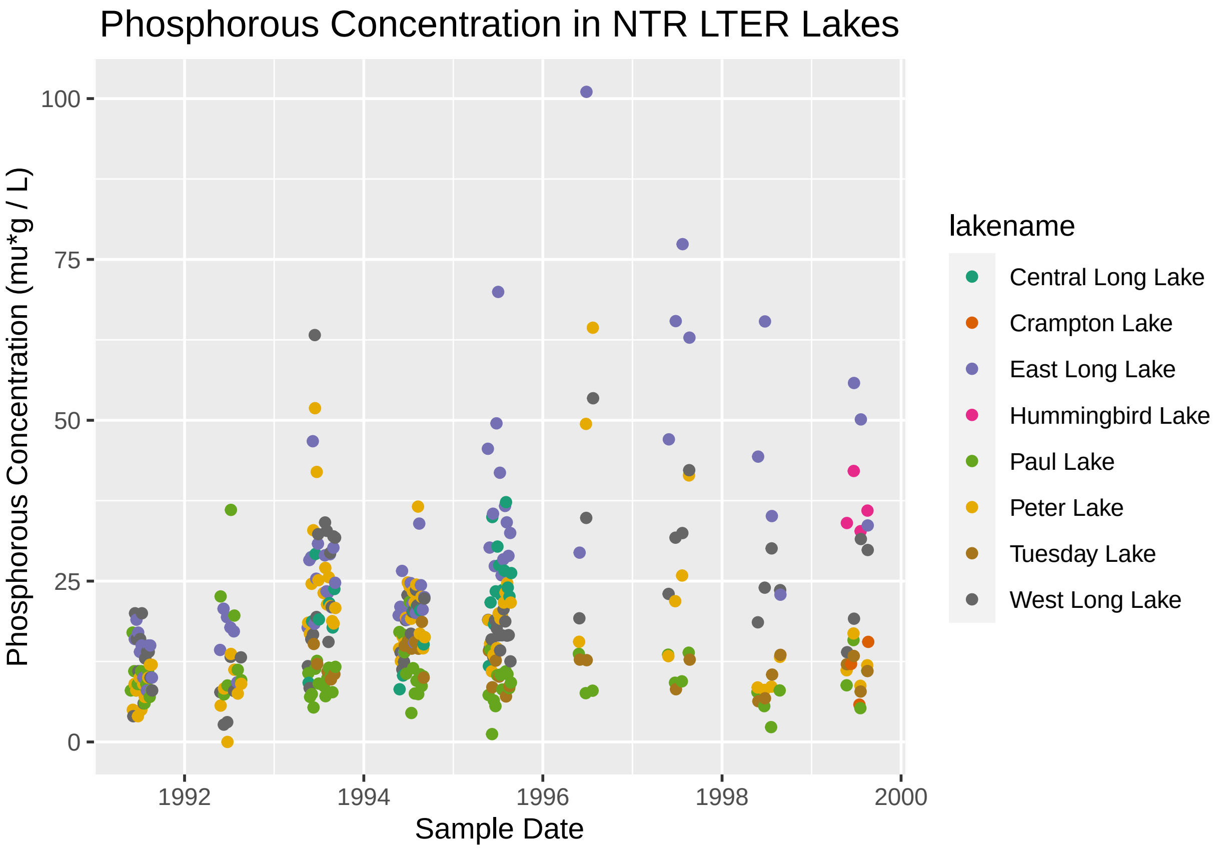


Figure 1: Figure 1: Phosphorous Concentrations in NTR LTER Lakes

Break page:

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?

The mean phosphorous concentrations across all lakes is 3.28626 mg / L and the mean nitrogen concentrations 610.9982 mg / L. The mean for nitrogen is much higher, and further literature review may be necessary

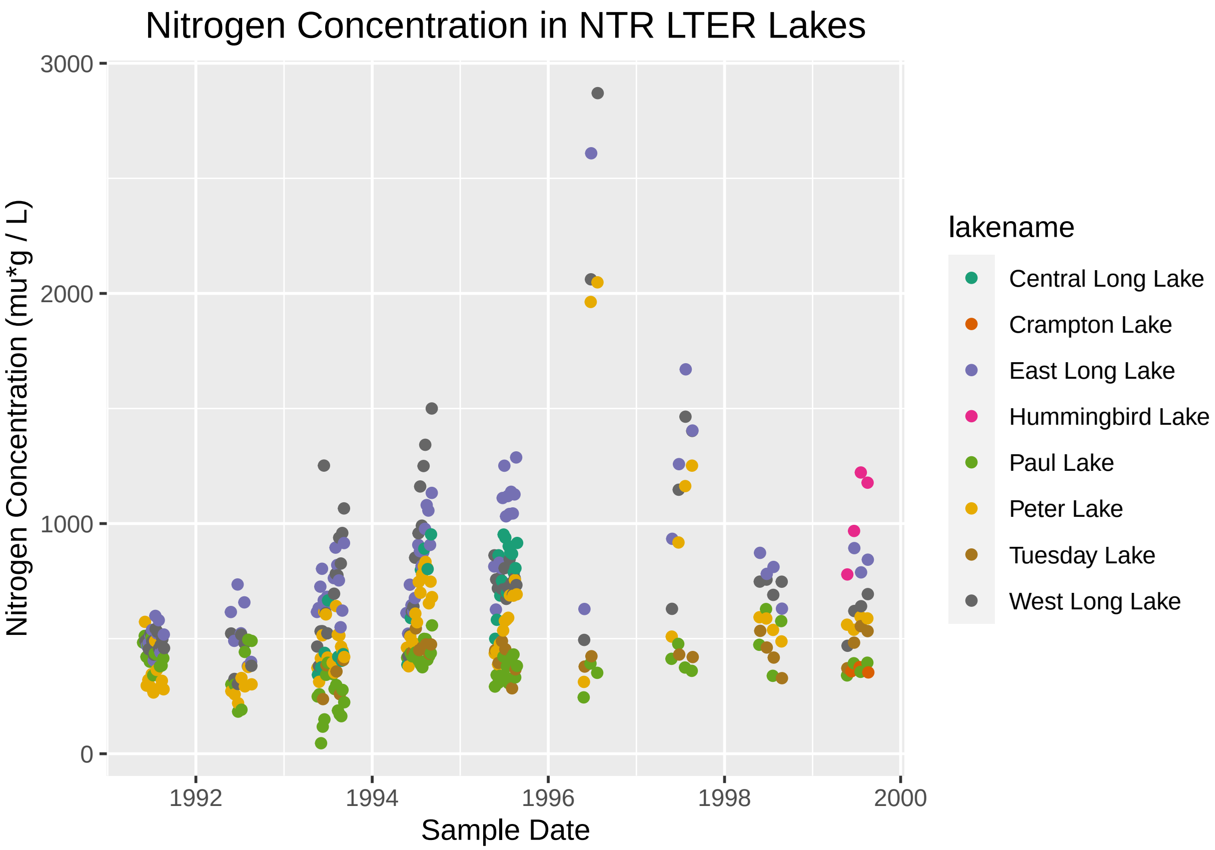


Figure 2: Figure 2: Nitrogen Concentrations in NTR LTER Lakes

to understand the safe levels of nitrogen and what lakes exceed this. The data above shows the trends in Nitrogen and Phosphorous concentrations in the NTR LTER lake sites. For many of the lakes, there appears to be relatively consistent phosphorous and nitrogen concentrations including for Paul Lake and Tuesday Lake. Prior to 1996, the concentrations are clustered among many of the lakes, however after 1996 particularly for the phosphorous concentrations, these clusters become more spread across the chart. Nitrogen in particular, appears to spike in 1996/97 and then return back to relatively similar levels, and the spike may be driving the higher mean concentration given that the minimum and maximum range from ~45 to >2000 mg/L.

Further statistical analysis will help to deepen the understanding of the nitrogen and phosphorous concentrations. In particular, a time series analysis could help to determine if there are trends over time in the lake concentrations. Analyzing trends in the data may bring to light important insights as to the health of the lakes, and potential turning points in phosphorous and nitrogen concentrations. Particular attention can be paid to the increase in nitrogen around 1996/97 and, if complemented by a literature review, potential causes of such a spike may be able to be identified. Regressions such as a GLM may also provide insight on the data, especially to find if there is a correlation between increases in phosphorous and nitrogen. Phosphorous also appears to increase near 1996/97 such that it may be a worthwhile analysis to compare the time series trends, as well as use GLMs to understand any correlations between the two nutrients.

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:*
```