11: Crafting Reports

Assignment completed by Halina Malinowski

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* - R markdown can use other languages including R, Python, and SQL.
- *Various outputs* - R markdown supports many types of outputs including, but not limited to PDF, HTML, dashboards, shiny apps, MS word, etc.

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).>

Reasons to use Markdown

- **Replicability** - ease of replicability for other researchers
- **Shareability** - ease of collaboration with others. R markdown and its connection to *Github* make it easy to work with other collaborators and edit code.
- **Clean Output** - R Markdown makes the final product in a clean output that is easy to read, understand, and reproduce. Figures and code can be added or removed within text very easily.
- **Clean/organized Code** - Using R Markdown helps create clean and organized code which is easier to understand and use in working with other people. Also makes it easier for you to revist your code and understand it. You can also select to only run certain chunks at a time which makes it easy to manage your code and output.
- **Integration with Github** - R Markdowns integration with Github gives it a nice platform for collaboration. The additional extension in R makes it really easy to push and pull and share code. *(I really love this feature! It was a little confusing at first, but I love how its essentially google docs for code!)*

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)

Table 1. Dataset descriptions

Dataset	Information
ECOTOX	Contains data about insects and the effects of neonicotinoids
EPAair	Contains data about air quality in North Carolina from 2018 - 2019
NEON_NIWO_Litter	Contains data about litter fall and woody debris collected at Niwot Ridge Long-Term Ecological Research (LTER) station
NTL-LTER_Lake	Contains data about nutrients as well as chemical and physical properties of several lakes in the North Temperate Lakes District in Wisconsin
NWIS_SiteFlowData	Contains data about river site flow in Nebraska

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

```
#subset data
colnames(NTL_LTER_Lake_Nutrients_Raw)
```

```
##  [1] "lakeid"     "lakename"   "year4"      "daynum"     "sampledate"
##  [6] "depth_id"   "depth"      "tn_ug"      "tp_ug"      "nh34"
## [11] "no23"       "po4"        "comments"
```

```
NTL_LTER_Subset <- NTL_LTER_Lake_Nutrients_Raw %>%
  select(lakename:sampledate, depth:po4)
colnames(NTL_LTER_Subset)
```

```
##  [1] "lakename"   "year4"      "daynum"     "sampledate" "depth"
##  [6] "tn_ug"      "tp_ug"      "nh34"       "no23"       "po4"
```

```
#filter for depth = 0
NTL_LTER_Subset <- NTL_LTER_Subset %>%
  filter(NTL_LTER_Subset$depth == 0)
unique(NTL_LTER_Subset$depth)
```

```
## [1] 0
```

```
#drop NAs
NTL_LTER_Subset <- NTL_LTER_Subset %>%
  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: Lake Nitrogen Concentration Summary Table

lakename	mean_nitrogen	min_nitrogen	max_nitrogen	sd_nitrogen
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: Lake Phosphorus Concentration Summary Table

lakename	mean_phosphorus	min_phosphorus	max_phosphorus	sd_phosphorus
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

lakename	mean_phosphorus	min_phosphorus	max_phosphorus	sd_phosphorus
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.

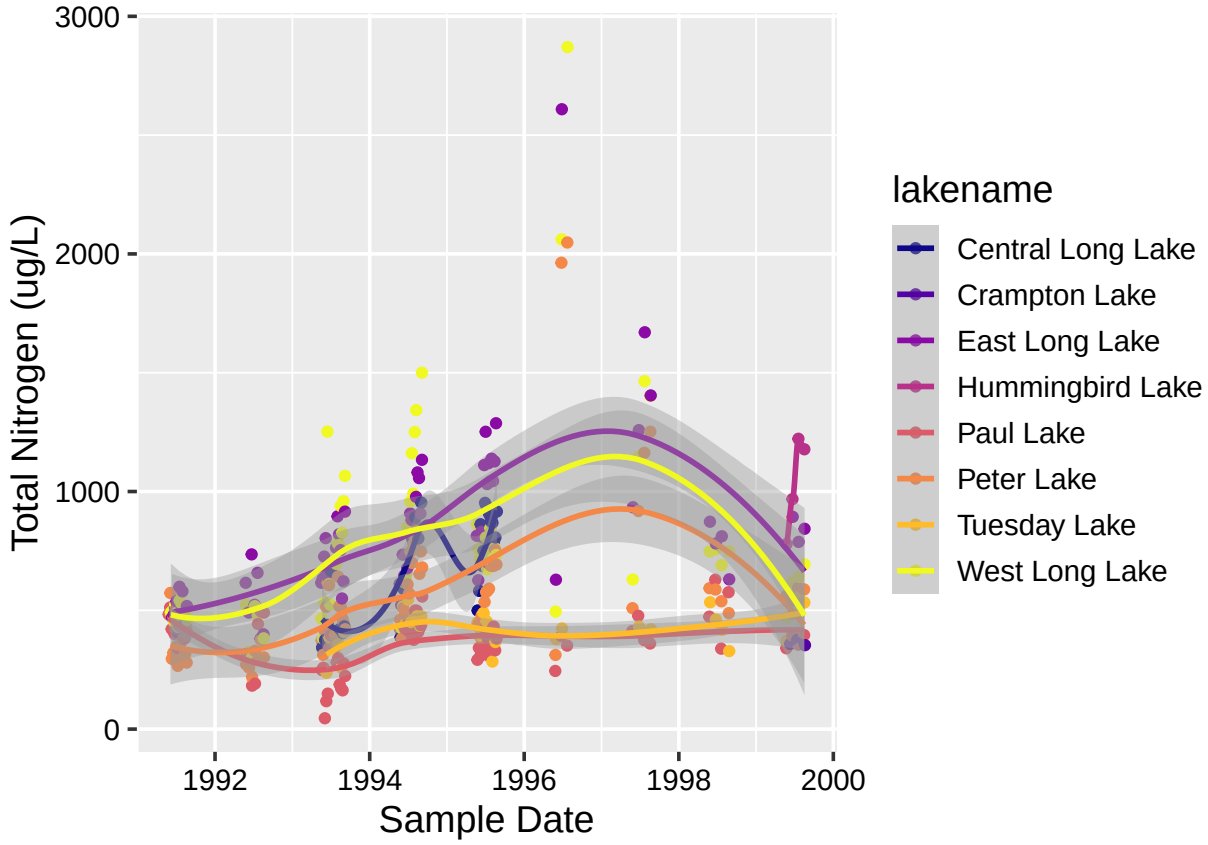

Figure 1: Total nitrogen over time

‘‘‘

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?

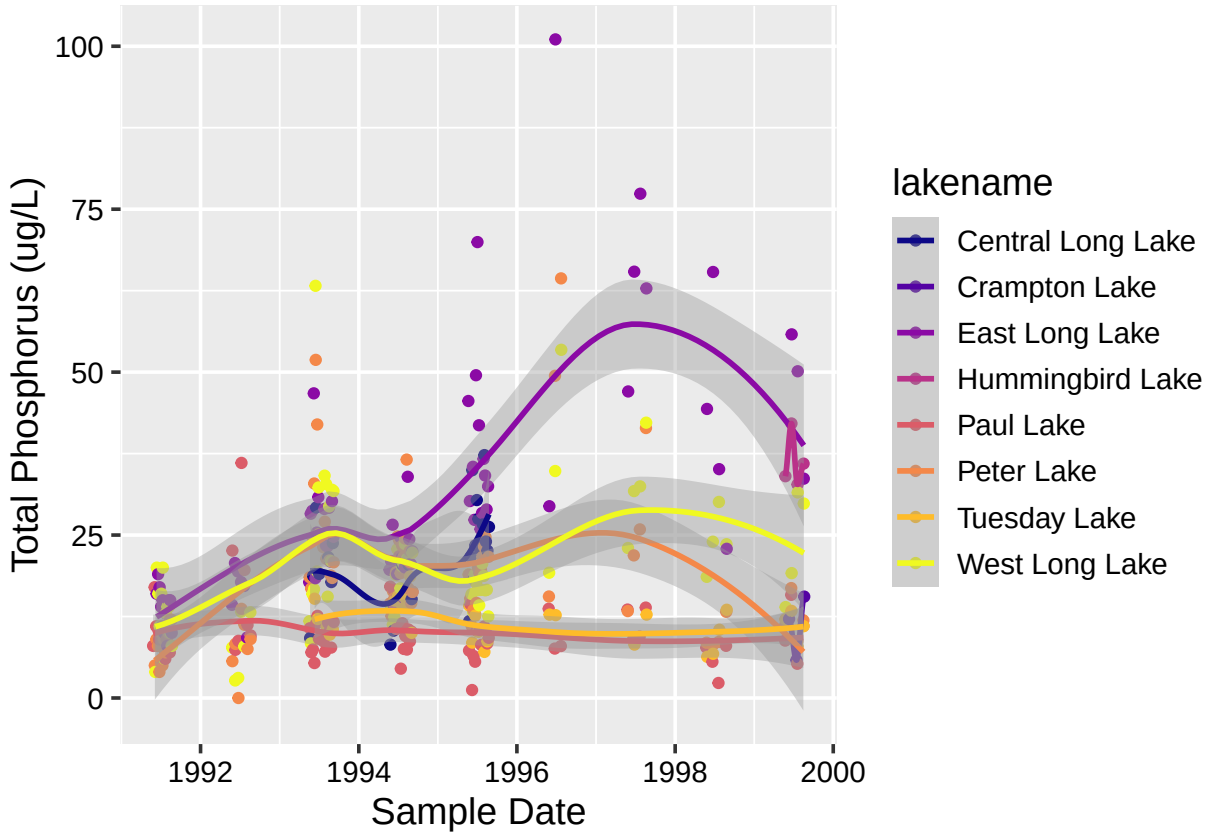

Figure 2: Total phosphorus over time

Results

After cleaning the data, two summary tables were created for nitrogen and phorphorus values of the NTL_LTER lakes. Overall nitrogen values were considerably higher than phosphorus values. From the summary tables it was clear that Hummingbird lake had the greatest mean total nitrogen and mean total phosphorus values. The above graphs displaying total nitrogen concentration over time and phosphorus concentration over time showed East Long Lake with the greatest values. From 1996 to 1998 many lakes saw increases in total nitrogen and phosphorus levels. West Long lake, Peter lake, and East Long lake all experienced large increase in nitrogen and phosphorus levels during this time. From the above results it appears nitrogen and phosphorus concentrations may be correlated in certain systems as they tend to increase and decrease in a similar pattern.

Further Analyses

To further explore this dataset and gain a better understanding of the trends and relationships present I would create generalized linear models to explore the relationship between nitrogen concentrations and environmental and spatial variables. I would conduct the same analyses for phosphorus levels. I would also determine if nitrogen and phosphorus levels are correlated. Additionally, I would explore the data with a time series analysis to see what types of trends their may be in the data and if nitrogen and phosphorus trends are similar. Perhaps we could analyze if there are any seasonal trends in nutrient concentrations or any other cyclical behaviors over longer time scales. A spatial analysis displaying the sites on a map and their proximity may also help orient the reader to understand if there are spatial variables that are effecting the lake nutrient concentrations. Further analyses are needed to gain a full understanding of this dataset and the changes and variability in nitrogen and phosphorus levels across the examined 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: