11: Crafting Reports

Environmental Data Analytics | John Fay & Luana Lima | Code Edited by Michael Gaffney

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 is a convenient way to produce **aesthetically pleasing reports**.
- R markdown operates through a **text language**, so it can be easy uploaded to github, unlike Microsoft products.
- In a realted sense, it can therefore also be **easily shared**, which creates *greater transparency* in your analysis.
- R Markdown can handle many **different coding languages**, like Python.
  - This can be useful for projects that are more easily accomplished in another language due to the availability of packages.

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)

File Name	Description
NWIS_SiteInfo_NE_RAW.csv	NWIS Site Information
NTL-LTER_Lake_Carbon_Raw.csv	NTL LTER Lake: Carbon 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] "/Users/michaelgaffney/Documents/Duke University/Nicholas School of the Environment/05 Spring 20
```

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

#set the theme
mytheme <- theme_dark() +
  theme(axis.text = element_text(color = "black"),
        legend.position = "left")
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.

```
##   lakeid  lakename year4 daynum sampledate depth_id depth tn_ug tp_ug nh34 no23
## 1      L Paul Lake  1991    140    5/20/91        1  0.00   538    25   NA   NA
## 2      L Paul Lake  1991    140    5/20/91        2  0.85   285    14   NA   NA
## 3      L Paul Lake  1991    140    5/20/91        3  1.75   399    14   NA   NA
## 4      L Paul Lake  1991    140    5/20/91        4  3.00   453    14   NA   NA
## 5      L Paul Lake  1991    140    5/20/91        5  4.00   363    13   NA   NA
## 6      L Paul Lake  1991    140    5/20/91        6  6.00   583    37   NA   NA
##   po4 comments
## 1  NA
## 2  NA
## 3  NA
## 4  NA
## 5  NA
## 6  NA
```

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

```
NTL.nutrients.final <- NTL.nutrients.data %>%
  select(-c("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: Surface Samples: Total Nitrogen

lakename	mean_tn_ug	min_tn_ug	max_tn_ug	sd_tn_ug
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: Surface Samples: Total Phospherous

lakename	mean_tp_ug	min_tp_ug	max_tp_ug	sd_tp_ug
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.

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?

In the above plots, I analyzed the nitrogen and phosphorous content of 8 different lakes over time, from 1991 to 2000. These 8 lakes are part of a long term ecological research program (LTER); learning about the phosphorous and nitrogen content for these lakes can tell us a great deal about the overall health of the lakes. In circumstances where lakes are affected by human development (either cities or agricultural production), they will often experience nutrient overloading, meaning that they will have more nutrients like phosphorous and nitrogen than they would in a natural state. This can lead to dramatically destabilized conditions in the lakes, like algal blooms that wipe out other wildlife. Looking at the nitrogen plot (figure 1), we can see that for East Long, West Long, and Peter Lakes, there appears to be a general trend from 1991 to 1997 of increasing nitrogen content, which then appears to decrease in the years afterwards. The other lakes with sufficient data appear to have substantially more stable nitrogen measurements over time. Meanwhile, turning to the phosphorous plot (figure 2), East Long lake appears to display a similar trend in increasing phosphorous from 1991 to 1997, which then begins to decline. West Long and Peter lake show a

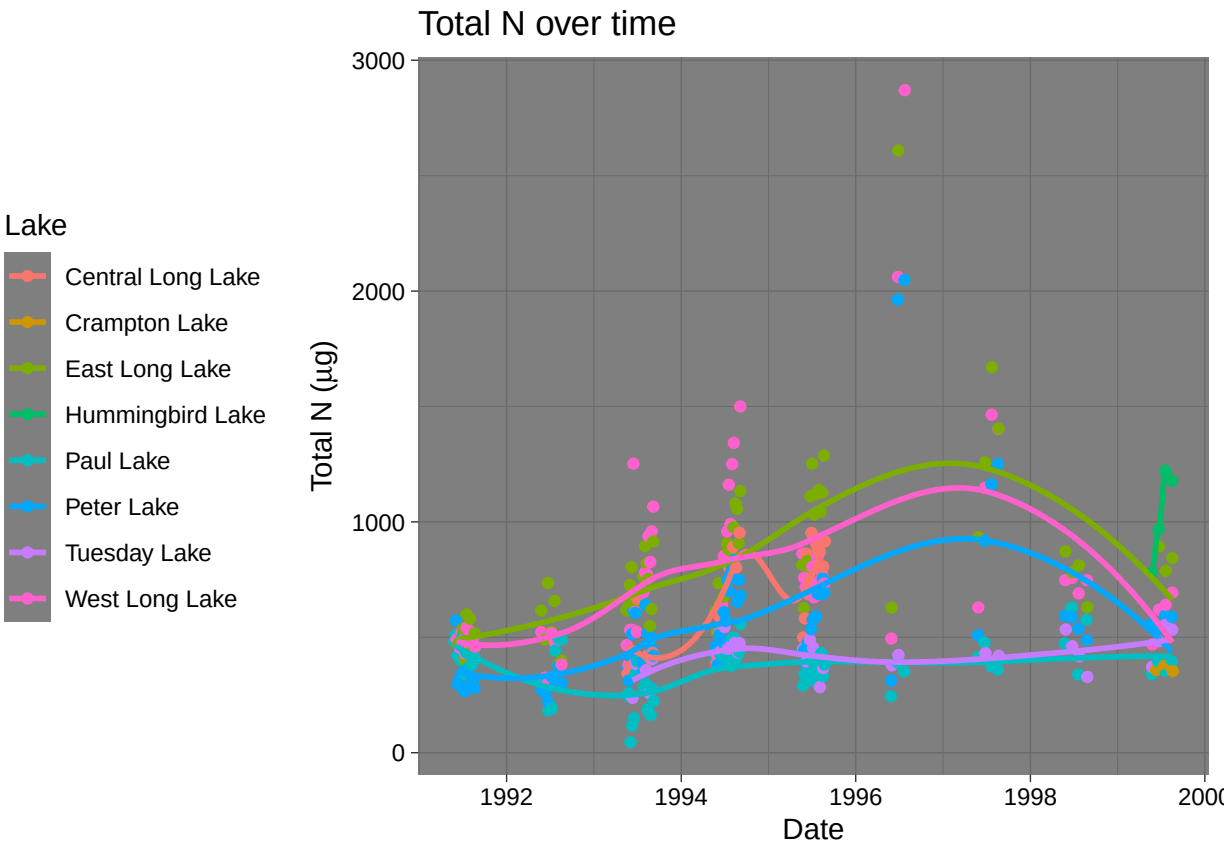

Figure 1: Total Nitrogen

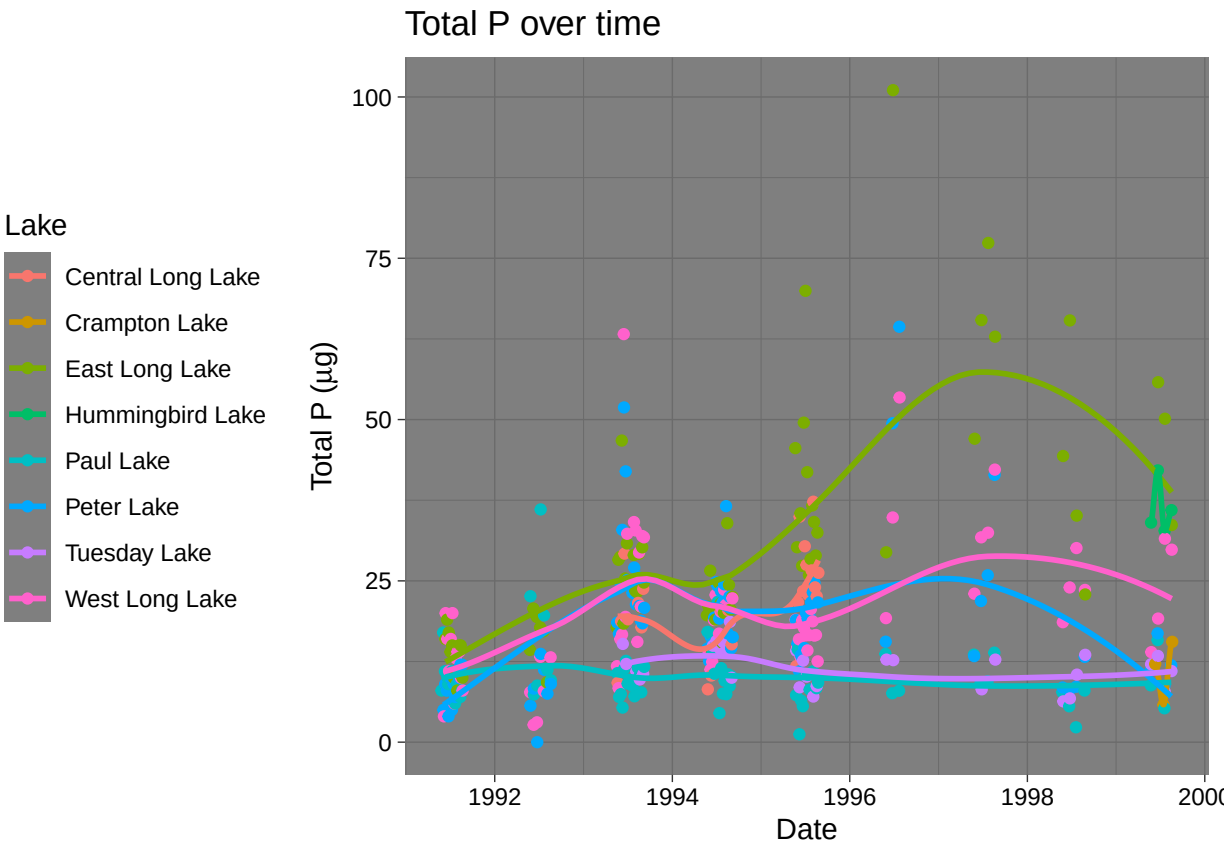

Figure 2: Total Phospherous

slightly increasing trend, but not nearly as substantial. The other lakes do not appear to show a substantial trend. We might conclude that the lakes showing substantial change also had nearby land use changes.

With further analysis, several different questions could be answered about these lakes. We could, for example, ask whether or not the trend that is visible is statistically significant, or whether it is largely generated by random processes. This could be accomplished using a time series analysis. We could also ask whether there is a relationship between nitrogen and phosphorous content among the different lakes, which would involve constructing a linear model. A similar model could compare nitrogen and phosphorous measurement means between different lakes. With additional data, more robust questions could be asked. For example, with land use data, we could determine what types of changes are associated with the increasing trends. With species data, we could also ask what effect the increased nitrogen and phosphorous load has had.

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: