11: Crafting Reports

Sara Diamond

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).>

- Easy to understand
- Good for reproducibility ** It is easy to compile reports **
- I like that we are able to knit to PDFs or HTML *

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 <- c("EPA Air", "NEON NWIO Litter",
                  "NTL LTER Lake Chem", "NTL LTER Lake Nutrients",
                  "Ozone", "Wind Speed")

dataset.info <- c("PM25 and O3 information",
                  "different litter information",
                  "lake chemistry info", "lake nutrients info",
                  "time series ozone", "data regarding wind speeds")

dataframe <- data.frame(dataset.name, dataset.info)

knitr::kable(dataframe, caption = "Table with Datasets and Information",
             col.names = c("Datasets", "Information"))
```

Table 1: Table with Datasets and Information

Datasets	Information
EPA Air	PM25 and O3 information
NEON NWIO Litter	different litter information
NTL LTER Lake Chem	lake chemistry info
NTL LTER Lake Nutrients	lake nutrients info
Ozone	time series ozone
Wind Speed	data regarding wind speeds

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.

```
#getting working directory
getwd()
```

```
## [1] "/Users/saradiamond/Documents/Environmental_Data_Analytics_2022"
```

```
setwd("/Users/saradiamond/Documents/Environmental_Data_Analytics_2022/Lessons")

#loading packages
library(tidyverse)
library(lubridate)
library(knitr)

#setting theme
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.

```
##   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
```

```
## [1] "Date"
```

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

```
#wrangling the data
NTL.LTER.Processed <-
  NTL.LTER.Nutrients %>%
  select(-c(lakeid, depth_id, comments)) %>%
  filter(depth == 0) %>%
    drop_na()
#saving processed file
write.csv(NTL.LTER.Processed,
          row.names = FALSE,
          file = "./Data/Processed/NTL-LTER_Processed.csv")
```

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

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

Table 3: Summary of Total Phosphorous

Lake Name	Mean	Max	Min	Std Dev
Central Long Lake	21.70981	37.270	8.190	7.076388
Crampton Lake	11.16033	15.555	5.803	4.946759
East Long Lake	29.28984	101.050	8.000	17.375710
Hummingbird Lake	36.21925	42.119	32.765	4.146717
Paul Lake	10.45606	36.070	1.222	4.805142
Peter Lake	18.39153	64.383	0.000	10.976205
Tuesday Lake	11.71853	18.663	6.325	3.044289
West Long Lake	19.82981	63.243	2.690	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.

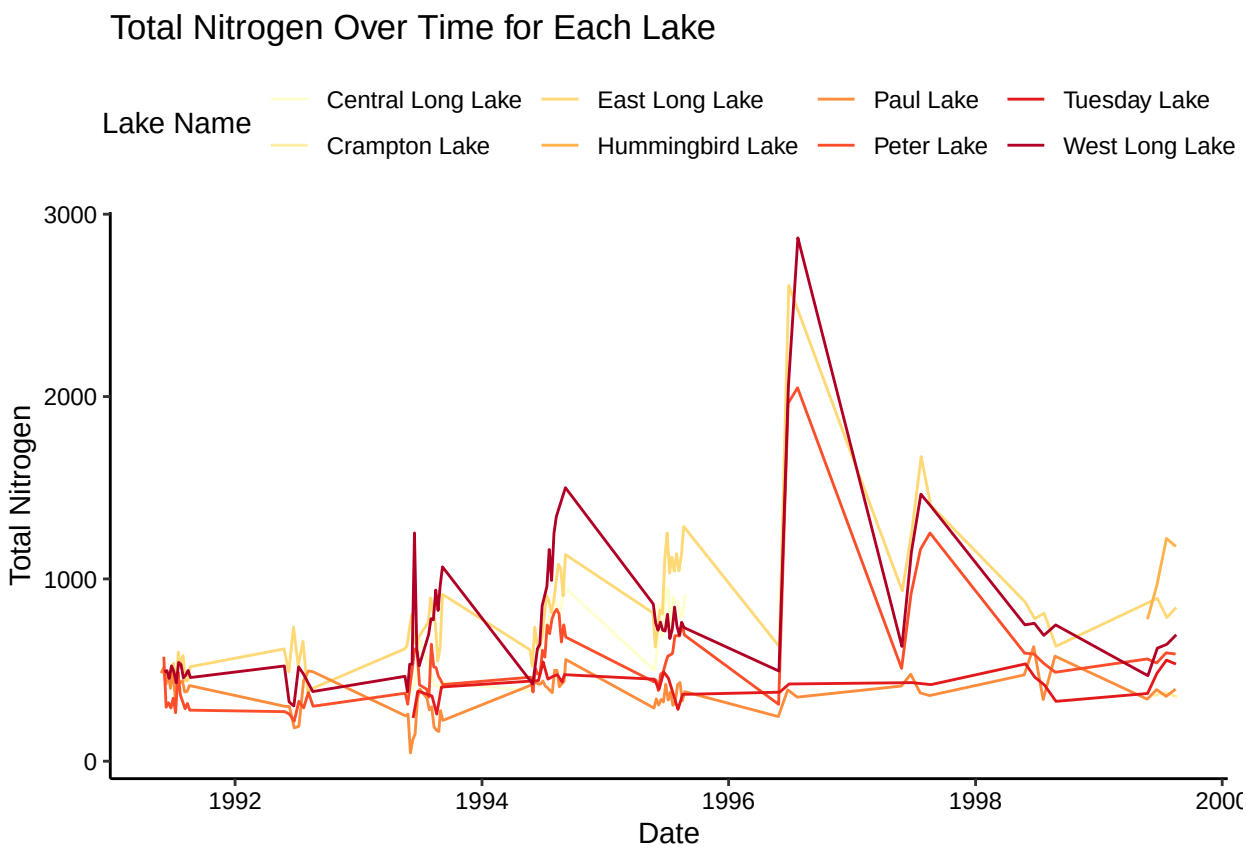

Figure 1: Total Nitrogen Over Time for Each Lake

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?

The data set we used in this analysis was looking at different lakes in the Northern Temperate Lakes District in Wisconsin, USA. The data we have looked at was collected by an ecological research station over time. Specifically, we were looking at the total amount of nitrogen and phosphorous nutrients within the various lakes. In our analysis we were able to determine different statistics (average, maximum, minimum, and standard deviation) for the two different nutrients and compare our findings of each. It was found that for total nitrogen, Hummingbird Lake had the highest amount out of all 8 lakes we were examining. Additionally, by looking at the plot of total nitrogen over time, we can see that many of the lakes fluctuated in their total nitrogen, but West Long Lake, Peter Lake and East Long Lake had much higher fluctuations in between 1996 and 1998. As for phosphorous, East Long Lake had the highest maximum, but Hummingbird Lake had the highest total average. When looking at the plot for total phosphorous over time it is clear that East Long Lake had the most variability but we would have to do further analysis to determine exactly why that is.

In order to determine more information, we might want to look to see if other variables are causeing an

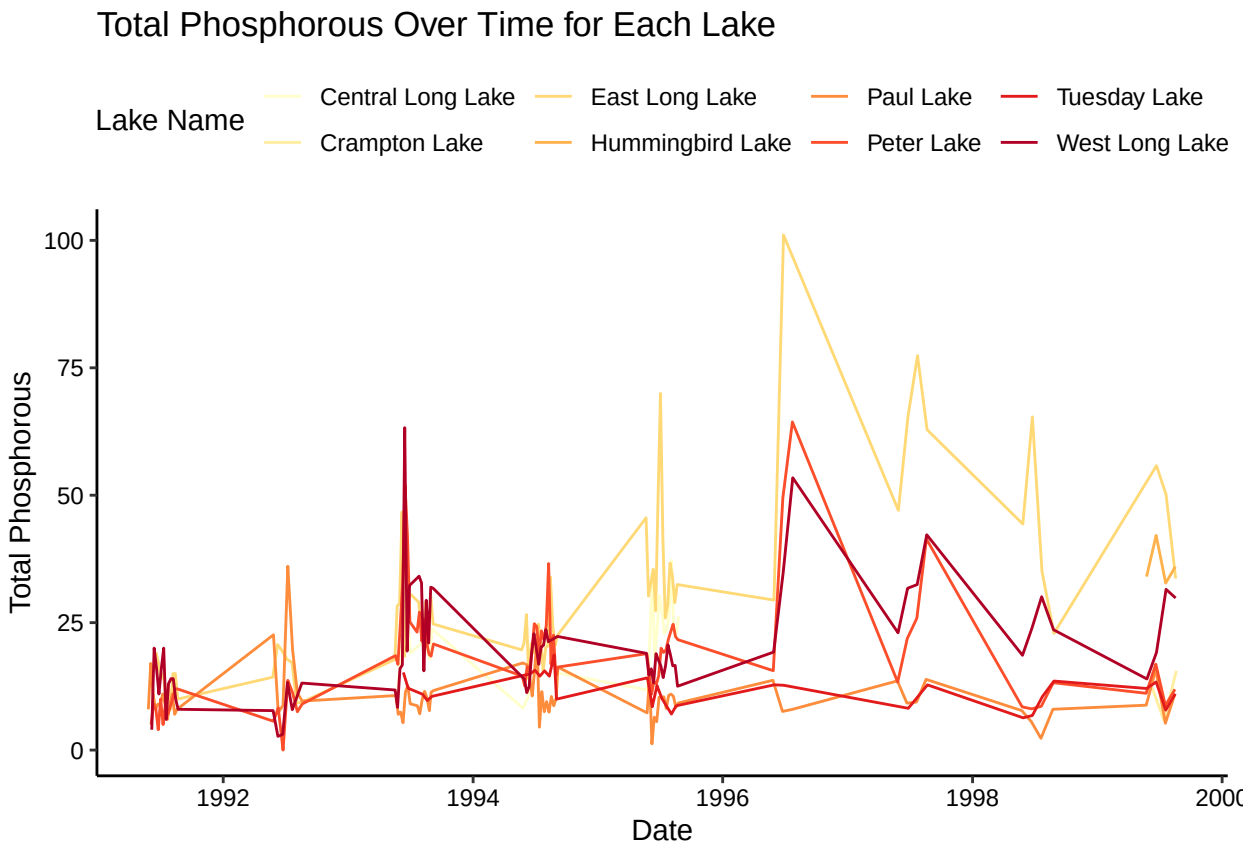

Figure 2: Total Phosphorous Over Time for Each Lake

increase or descrease in the total amount of nutrients over time. We could do this by running some simple linear models to see if any variables are dependent on the total amount of nitrogen or phosphorous. For example, we could look at depth to see if this played a role or we could look at the presence of other nutrients in the total nitrogen/phosphorous observed. By conducting these analyses we could accept or reject a hypothesis saying that the presence of another nutrient had an effect on the total nitrogen or phosphorous amount in each lake.

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: