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...

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 allows for the use of many different languages including R, Python, and SQL. * R Markdown uses a productive interface to weave together narrative text and code to produce elegantly formatted output. * R Markdown also supports dozens of dynamic and static output formats including HTML, PDF, MS Wods, Beamer, HTML5 slides, and more.

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	Details
EPAair_O3_NC2019_raw.csv	Contains ozone concentrations in North Carolina in 2019.
EPAair_PM25_NC2019_raw.csv	Contains PM 2.5 concentrations in North Carolina in 2019.
NTL-LTER_Lake_Nutrients_Raw.csv	Contains data from studies on several lakes in the North Temperate Lakes District in Wisconsin, USA. Data were collected as part of the Long Term Ecological Research station established by the National Science Foundation.
ECOTOX_Neonicotinoids_Insects_raw.csv	Contains data from studies on several neonicotinoids and their effects insects.

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')
#`eval = FALSE` - code is not run in the knitted file
```

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/meganlundequam/Desktop/Spring 2022/Environmental Data Analytics/Git/Environmental_Data_A
```

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


mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "bottom")
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

```
NTL.nutrient.data.processed <-
  NTL.nutrient.data %>%
  select(lakename:sampledate, depth:po4) %>%
  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 Statistics for Nitrogen Concentrations in NC Lakes

lakename	meannit	maxnit	minnit	sdnit
Bergner Lake	471.3840	626.5504	360.5784	92.52036
Bolger Bog	800.5791	1334.3991	647.7846	197.59391
Brown Lake	667.4650	1094.6642	390.8921	185.81284
Central Long Lake	794.4133	2474.3030	157.1900	510.04678
Crampton Lake	351.9243	956.4060	163.3900	137.38049
Cranberry Bog	414.4075	494.5169	355.2214	47.42169
East Long Lake	848.9101	3316.8920	0.0000	492.11923
Hummingbird Lake	915.1903	1462.5070	612.6930	200.34164
Inkpot Lake	464.0169	549.1784	390.2457	57.29937

lakename	meannit	maxnit	minnit	sdnit
Morris Lake	639.8115	767.4801	545.4971	80.28057
North Gate Bog	498.4990	589.2487	412.3507	50.09471
Paul Lake	433.3314	2099.0000	45.6700	308.23787
Peter Lake	534.3640	3497.6990	111.2500	400.92843
Plum Lake	392.4660	447.4974	324.6816	45.37608
Raspberry Lake	394.4905	426.0130	368.8612	20.33686
Reddington Lake	668.8188	790.9104	583.0434	67.51347
Roach Lake	253.6822	287.1464	229.4159	17.08657
Tender Bog	545.2030	587.6459	504.5756	42.13848
Tenderfoot Lake	461.6497	615.7022	359.4719	80.55970
Tuesday Lake	532.9443	1572.2620	215.4970	211.69369
Ward Lake	488.7789	658.2269	365.1683	73.22381
West Long Lake	753.3605	2950.3430	155.6100	489.35476

Table 3: Summary Statistics for Phosphorus Concentrations in NC Lakes

lakename	meanphos	maxphos	minphos	sdphos
Bergner Lake	12.247609	23.54200	8.3116307	3.619843
Bog Pot	42.700000	59.09800	29.2370000	13.564224
Bolger Bog	52.317400	154.77538	26.2790000	32.941216
Brown Lake	41.247157	80.70900	22.6295107	17.737526
Central Long Lake	27.650641	86.76200	3.8700000	15.055482
Crampton Lake	11.519615	53.33300	3.2700000	8.873020
Cranberry Bog	13.793455	28.54400	8.3289575	5.101257
East Long Lake	58.782476	352.05600	0.0000000	62.109048
Eds Bog	36.614500	40.75000	33.7970000	3.293310
Forest Service Bog	17.532250	23.15600	13.2480000	4.169864
Hummingbird Lake	34.376353	69.86900	17.8107015	10.788058
Inkpot Lake	21.656568	37.88900	11.8956886	7.433935
Kickapoo Lake	32.879250	37.20000	25.9960000	4.857109
Morris Lake	24.331630	34.51800	18.3060522	4.835955
North Gate Bog	15.723024	30.81200	10.1473972	5.288118
Paul Lake	16.097701	123.29000	-6.3488500	16.711673
Peter Lake	25.209163	157.25000	-3.0970000	23.316691
Plum Lake	18.430711	35.98800	8.4414848	7.037070
Raspberry Lake	13.220990	20.48700	9.8049848	2.965015
Reddington Lake	18.677777	35.10200	11.1614607	6.761786
Roach Lake	9.948343	17.82526	4.8086825	3.652945
Tender Bog	12.722948	16.67473	9.7924446	3.290124
Tenderfoot Lake	28.294714	90.21764	5.1512240	20.363216
Tuesday Lake	20.310104	101.89000	-0.7783225	12.926009
Ward Lake	25.048789	35.12000	18.9394796	4.610931
West Long Lake	37.010093	251.67400	2.6900000	39.090908

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.

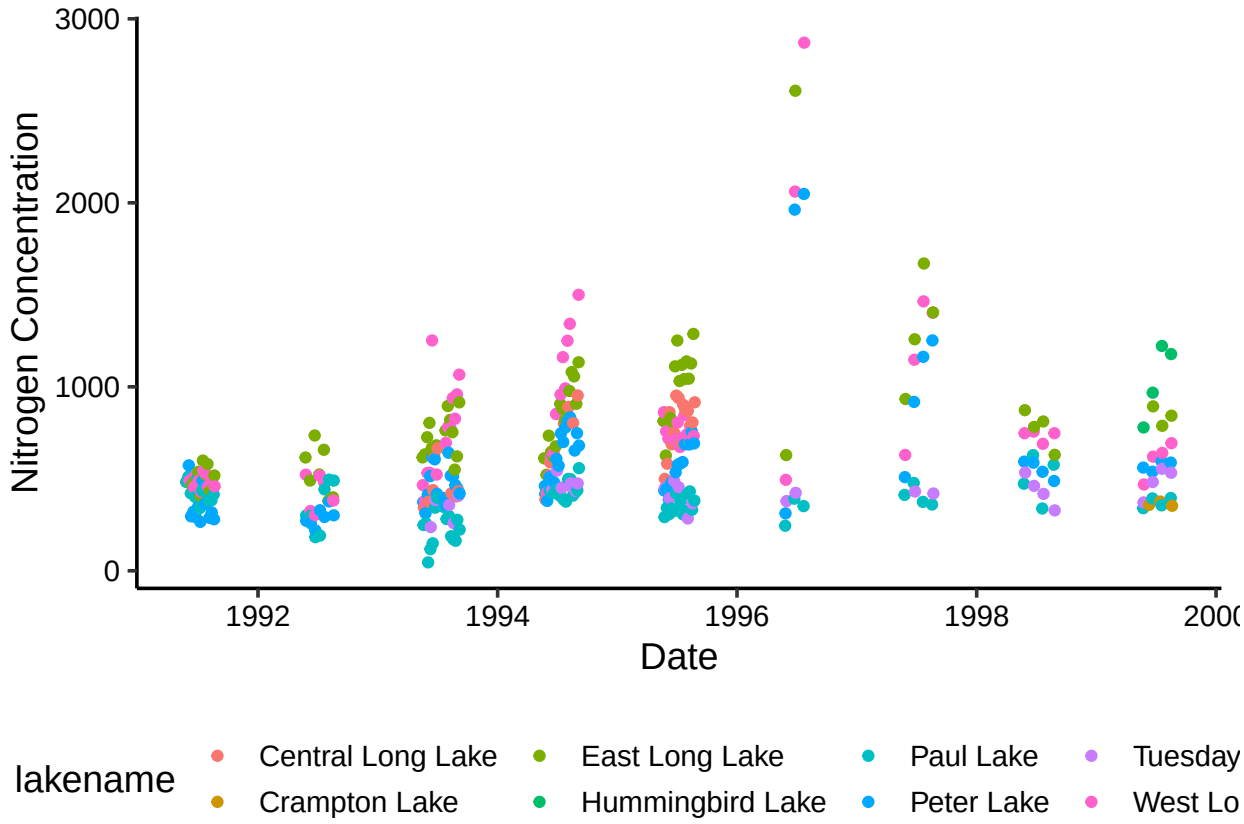

Figure 1: Total Nitrogen Over Time

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?

Based on the above data wrangling and plots, nitrogen and phosphorus levels appear to have been fluctuating at different rates in different locations across time. From 1991 to 1995, both nitrogen and phosphorus concentrations appear to have been generally increasing across the lakes sampled. After 1996, concentrations of both nutrients generally decreased, but the trend is slightly less pronounced than the increasing trend observed before 1996. East Lake Lake appears to consistently have the highest levels of nitrogen and phosphorus among the sample population, but appears to follow an increasing trend until 1996 where concentrations appeared to decrease. Concentrations in Paul Lake appear to be consistently low compared to the other lakes in the sample population.

To further analyze the trends observed above, an important first step would be to analyze the concentration over time for each lake individually to better visualize trends. We could also look at the concentration fluctuations at different sample depths by applying a continuous gradient filter to depth for each lake. We

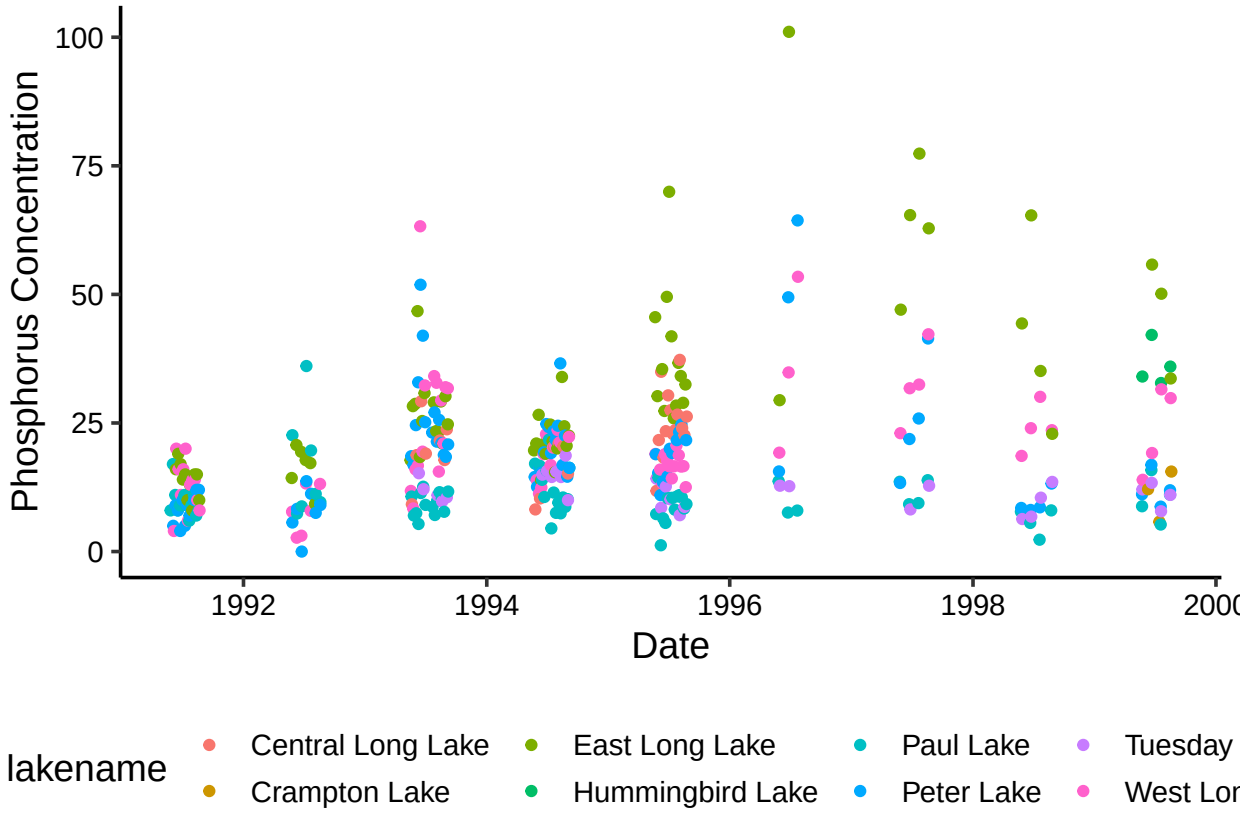

Figure 2: Total Phosphorus Over Time

could also look at the trends of other nutrients in the dataset to see if they follow a similar trend or if changes could be in relation to one another. If we wanted to understand the concentration trends even better, we could investigate the impact of potential sources of nitrogen and phosphorus and how those might have changed over time. To begin that investigation using this dataset and a bit of additional data, we could use spatial analysis to see where the lakes are relative to one another, and where they are relative to potential sources of nitrogen and phosphorus (i.e., residential areas).

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: