11: Crafting Reports

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

Nancy Bao

Spring 2021

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

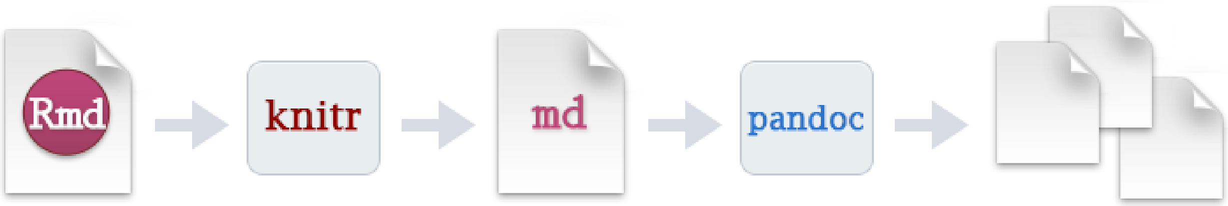
RESOURCES

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).>

- ***concise in keeping code and report together***
- ***easy for creating data visualization***
- ***generates professional deliverable***

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. Data set Descriptions

Data Set	Description of data set
CDC Social Vulnerability Index	2018 North Carolina county level data collected from the CDC Agency for Toxic Substances and Disease Registry
ECOTOX Neonicotinoids	Neonicotinoids and insect effects data collected from the US EPA ECOTOX Knowledgebase
EPA Air Quality	Air quality data from EPA monitoring sites in North Carolina measuring PM2.5 and ozone from 2017 to 2018
NEON Niwot Ridge litter	Small woody debris and litter data collected from 2016 to 2019 at the Niwot Ridge Long-Term Ecological Research (LTER) station
NTL-LTER Lake Datasets	Data collected from lakes in the North Temperate Lakes District in Wisconsin, USA from 1984 to 2016
USGS Streamflow data for site 02085000	Streamflow data from the Eno River, NC streamflow gage site 02085000 in North Carolina collected from 1928-01-01 and 2019-12-26.

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.

```
#Check working directory
getwd()
```

```
## [1] "/Users/Nancy/Desktop/Semester 4/ENV 872L/Environmental_Data_Analytics_2021"
```

```
#Load libraries
library(tidyverse)
library(lubridate)
library(knitr)
library(RColorBrewer) #loaded for more color palettes for the plots-used Dark2 palette
library(kableExtra)
#I used the URL below to read about kableExtra for:
#formatting my tables for the Data Exploration, Wrangling, Visualization section
#https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html
#Set a ggplot theme
A11_theme <-theme_bw(base_size=14)+
            theme(legend.position = "top",
            legend.justification = "center",
            legend.text = element_text(size = 14,color = "black"),
            legend.title = element_text(size = 14,color = "black",
                                        face= "bold"),
            plot.title = element_text(hjust = 0.5,size=14))
theme_set(A11_theme)
```

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 processed dataset for NTL-LTER_Lake_Nutrients_Raw dataset
NTL_nutrients_processed <- NTL_nutrients %>%
                            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: Descriptive Statistics for Total Nitrogen Concentrations in the Lakes at the North Temperate Lakes Long Term Ecological Research Site

Lake name	Mean (ug/L)	Standard Deviation (ug/L)	Minimum (ug/L)	Maximum (ug/L)
Central Long Lake	690.0469	209.09341	343.020	953.063
Crampton Lake	362.6813	12.05748	353.380	376.304
East Long Lake	810.7834	335.41457	380.620	2608.956
Hummingbird Lake	1036.6695	204.36889	779.053	1221.960
Paul Lake	368.7564	106.34741	45.670	628.625
Peter Lake	561.8752	305.64909	219.720	2048.151
Tuesday Lake	423.5605	78.84522	237.363	554.418
West Long Lake	762.6017	402.95992	303.170	2870.302

Table 3: Descriptive Statistics for Total Phosphorus Concentrations in the Lakes at the North Temperate Lakes Long Term Ecological Research Site

Lake name	Mean (ug/L)	Standard Deviation (ug/L)	Minimum (ug/L)	Maximum (ug/L)
Central Long Lake	21.70981	7.076388	8.190	37.270
Crampton Lake	11.16033	4.946759	5.803	15.555
East Long Lake	29.28984	17.375710	8.000	101.050
Hummingbird Lake	36.21925	4.146717	32.765	42.119
Paul Lake	10.45606	4.805142	1.222	36.070
Peter Lake	18.39153	10.976205	0.000	64.383
Tuesday Lake	11.71853	3.044289	6.325	18.663
West Long Lake	19.82981	10.541276	2.690	63.243

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.

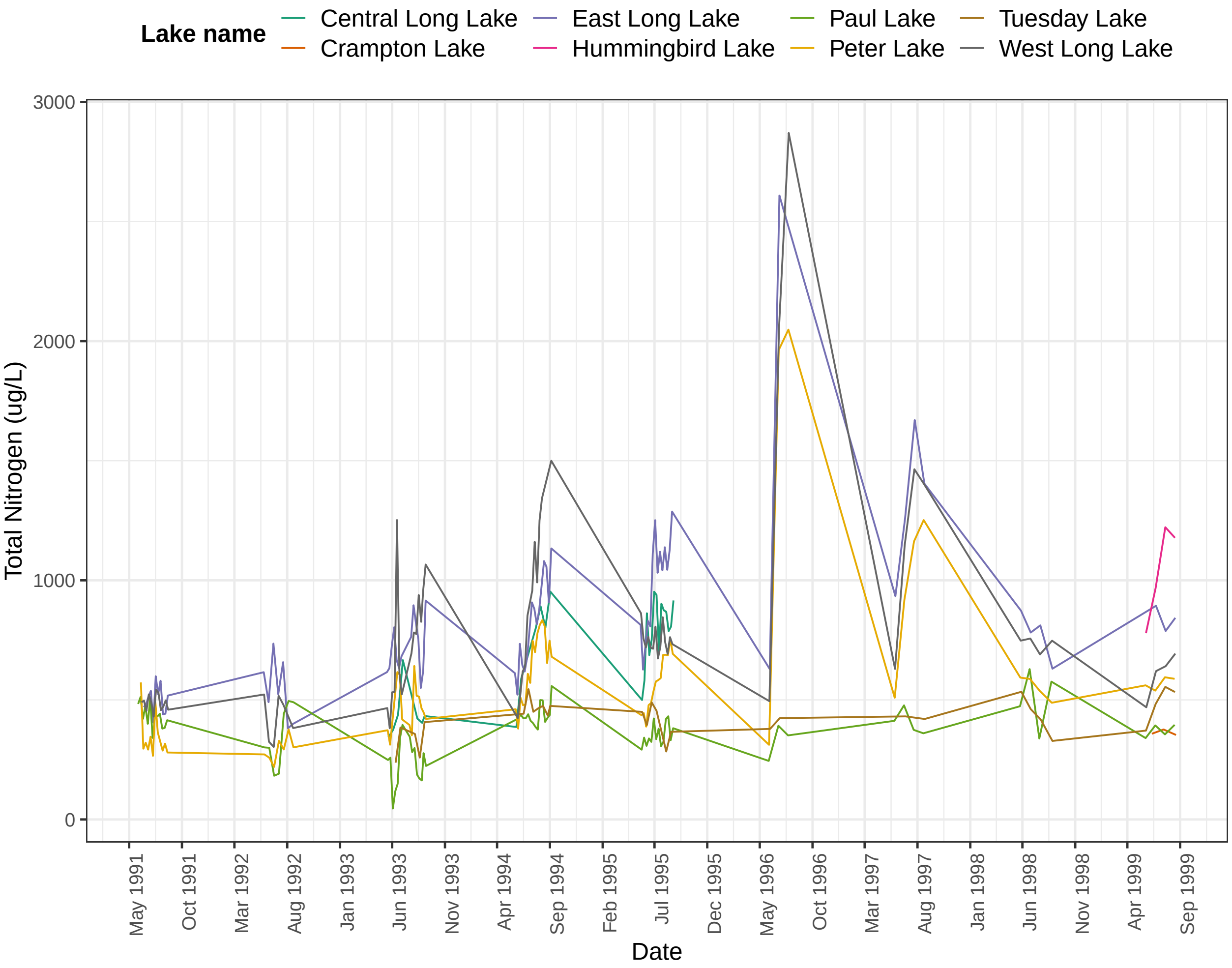


Figure 1: Total nitrogen surface (depth=0m) concentrations in lakes at the North Temperate Lakes Long Term Ecological Research Site

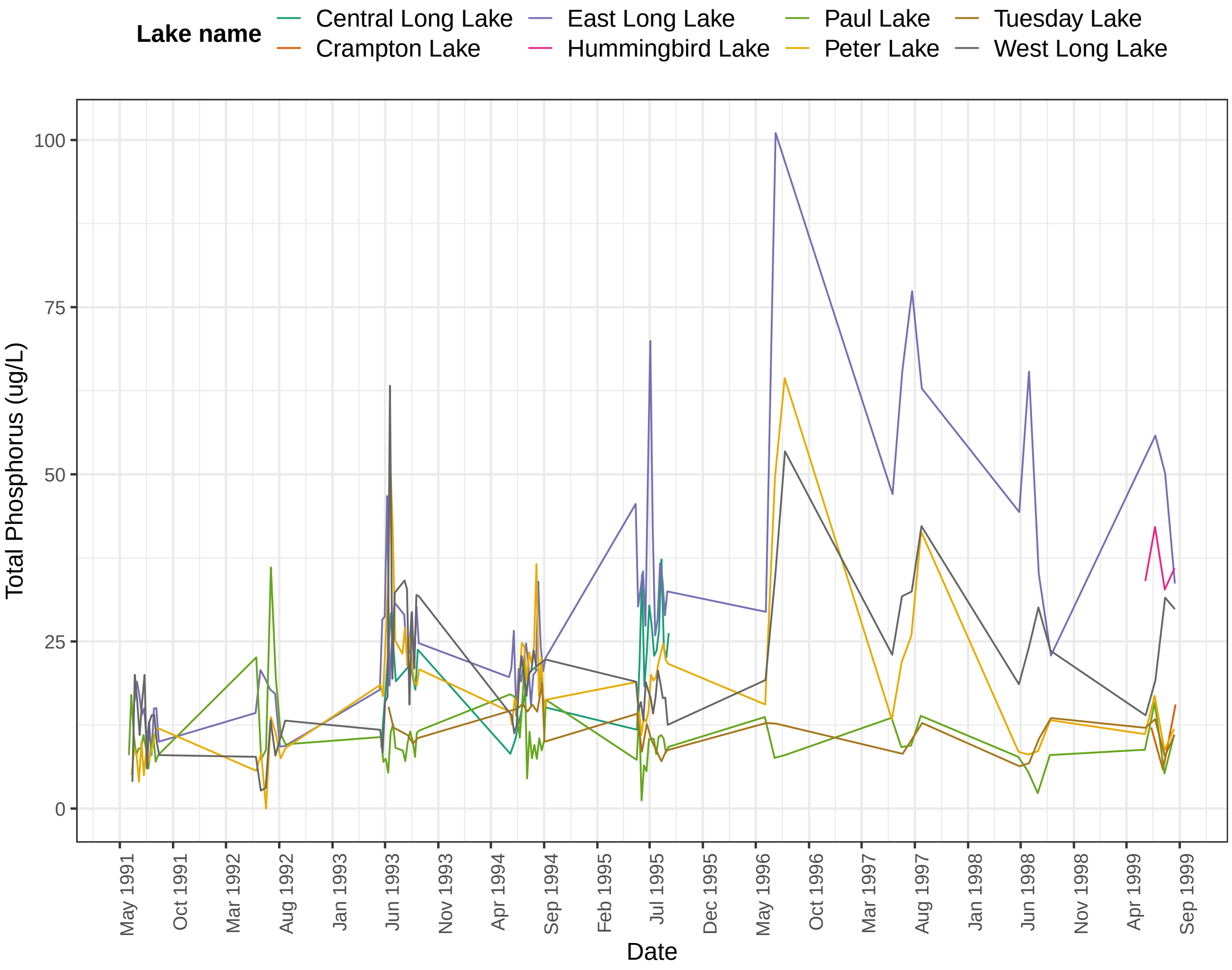


Figure 2: Total phosphorus surface (depth=0m) concentrations in lakes at the North Temperate Lakes Long Term Ecological Research Site

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?

> As shown in Figure 1 and 2 respectively, total nitrogen and total phosphorus concentrations were measured and collected across eight lakes from the North Temperate Lakes Long Term Ecological Research Site in Wisconsin, USA from May 1991 to August 1999. The following lakes were measured for their total nitrogen and total phosphorus nutrient concentrations in micrograms per liter: Central Long Lake, Crampton Lake, East Long Lake, Hummingbird Lake, Paul Lake, Peter Lake, Tuesday Lake, and West Long Lake.These nutrient concentrations (ug/L) are measured from the surface depth of each lake (depth=0m). In figure 1, the total nitrogen concentrations follow a seasonal pattern where concentrations peak during the late spring to mid-summer months and decrease during the late fall to winter months. In figure 2, the total phosphorus concentrations also a follow a seasonal pattern, where phosphorus concentrations peak in the summer to fall months. Of the eight lakes,total phosphorus and total nitrogen concentration measurements for

> East Long Lake across 1991 to 1999 are greater than that of the other lakes (Figs. 1 and 2). The nutrient concentrations for Peter Lake and Tuesday Lake do not greatly fluctuate from 1991 to 1999, but nutrient concentrations increased from May 1996 to March 1997 and decreased from May 1997 to August 1998 for East Long Lake, Peter Lake, and West Long Lake (Figs. 1 and 2). For the overall averages for total nitrogen concentration, Hummingbird Lake had the highest value at 1037 ug/L (Table 2) and for the overall averages for total phosphorus concentration, Hummingbird Lake had the highest value at 36.2 ug/L (Table 3).
>
> ---
>
> This dataset can be used to assess the impacts of nutrient pollution from upstream nonpoint runoff sources such as agricultural fertilizer and pesticide runoff surrounding Wisconsin from 1991 to 1999. The dataset can be used to explore research questions such as: "How the total phosphorus and nitrogen concentrations vary within a lake and between other lakes?" and "What factors change nitrogen and phosphorus concentrations in these lakes?" To elucidate the legacy effects of the nutrient effluent time series analyses can be conducted on each lake to assess trends across 1991 and 1999 and multiple linear regression analyses can be conducted to assess factors that are associated with total nitrogen and phosphorus lake concentrations. Furthermore, nitrogen nutrient concentrations can be further divided into the concentrations of nitrate versus ammonium to compare distribution of nitrogen forms in each lake. Others may be interested in looking at nutrient concentrations across varying depths in each of the lakes, in which including additional data such as water depths and other aspects such as seasonality may be included for further analyses on the effects of nutrient pollution within a lake and across different 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: