11: Crafting Reports

Blair Johnson

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

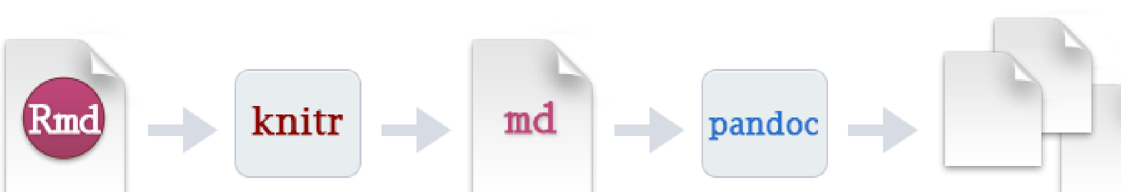
RESOURCES

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

R markdown then goes to base markdown

WHAT ELSE CAN R MARKDOWN DO?

See: https://rmarkdown.rstudio.com and class recording. * Languages other than R… * Various outputs…

WHY R MARKDOWN?

- R Markdown allows one to have a *centralized* point of access for data by having **text** and **code** in one place
- Good for *reproducibility* because one can easily adjust and re-run code to get new results and by linking code with output. Therefore, if an analysis changes, one can easily re-run the data in R.

- Can *integrate* data and visualizations easily through knitting through *multiple* output formats such as an **html**, **pdf**, or **word document**.

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.column1 <- c("EcoTox-Neonicotinoids-Insects", "EPAair_O3", "EPAair_PM25", "NEON_NIWO_Litter", "N
dataset.column1 <-as.data.frame(dataset.column1)

dataset.column2 <- c("This dataset shows data from studies on the effects of neonicotinoids (an insectic
dataset.column2<- as.data.frame(dataset.column2)

EDA.datasets <-cbind(dataset.column1, dataset.column2)
names(EDA.datasets) <-c("Dataset", "Data Description")

knitr::kable(EDA.datasets,
             caption = "EDA Course Datasets", table.attri ="style='width:30%;")
```

Table 1: EDA Course Datasets

Dataset	Data Description
EcoTox-Neonicotinoids-Insects	This dataset shows data from studies on the effects of neonicotinoids (an insecticide) on insects
EPAair_O3	This dataset shows ozone concentrations in North Carolina in 2018 and 2019
EPAair_PM25	This dataset shows particulate matter (2.5) concentrations in North Carolina in 2018 and 2019
NEON_NIWO_Litter	This dataset shows trap data from 2016-2019 of litter and woody debris at the Niwot Ridge Long-Term Ecological Research Station
NTL-LTER_Lake	This dataset shows physical, chemical, and nutrient data from lakes in the North Temperate Lakes District in Wisconsin

Dataset	Data Description
SVI2018-NORTHCAROLINA-county	This dataset is a shapefile that shows social vulnerability data for North Carolina on the county level
USGS_Site02085000-Flow	This datset shows streamflow data at USGS gage site 02085000 (Eno River in Hillsborough)

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

```
NTL.data.processed <- NTL.data %>%
  select(lakename:sampledate, depth:po4) %>%
  filter(depth == 0) %>%
  filter(!is.na(tn_ug) & !is.na(tp_ug) & !is.na(nh34) & !is.na(no23) & !is.na(po4))
```

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 Lakes

lakename	mean.totalnitrogen	min.totalnitrogen	max.totalnitrogen	sd.totalnitrogen
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

lakename	mean.totalnitrogen	min.totalnitrogen	max.totalnitrogen	sd.totalnitrogen
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: Summary Statistics for Phosphorus Concentrations in Lakes

lakename	mean.totalphosphorus	min.totalphosphorus	max.totalphosphorus	sd.totalphosphorus
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

The figures above illustrate the total concentrations of nitrogen and phosphorus over time (1991-1999) across 8 lakes in Wisconsin. Based on the visualizations, East Long Lake, Peter Lake, and West Long Lake display a sharp increase in total phosphorus and total nitrogen between 1996 and 1997. West Long Lake and East Long Lake also show sharp increases in phosphorus between 1993-1994 and 1995-1996 which lead to phosphorus concentrations that are much larger than those in the other 6 lakes during that time period. In addition, Crampton Lake, Tuesday Lake, Paul lake show less variability in the phosphorus and nitrogen concentrations over time because the visualizations display less peaks. These lakes also have the smaller standard deviations for total nitrogen and phosphorus relative to the other lakes.

Some appropriate next steps for analyzing this dataset would be to determine (or make assumptions) as to why the nutrient concentrations in each lake vary over time and why some lakes such as East Long Lake, Peter Lake, and West Long Lake may experience spikes in total phosphorus and total nitrogen. This could be done by assessing the concentrations of additional nutrient concentrations in the raw dataset (ex. nh34, n023) to gain insight as to what human and/or biological activities contribute to the prevalance of certain nutrients in the lake. These concentrations can also be plotted over time using the geom_line function to better understand trends.

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.

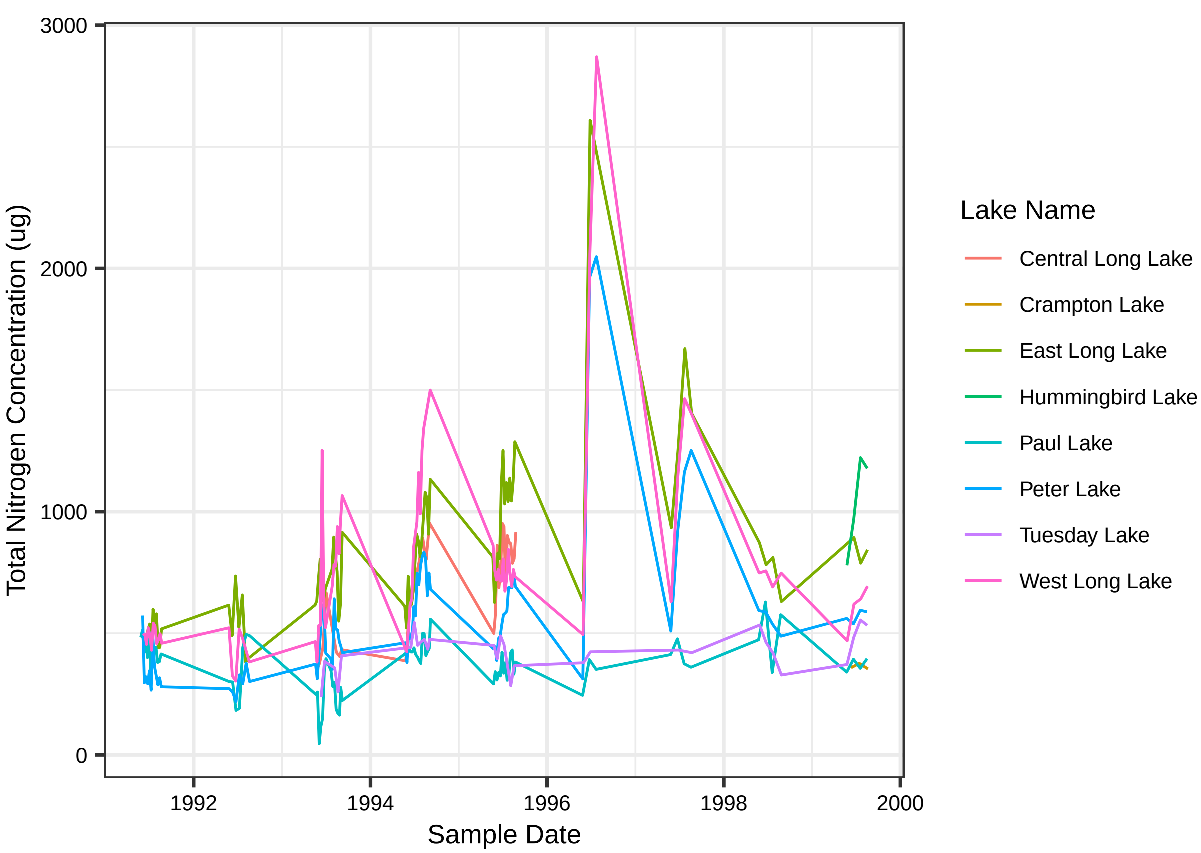


Figure 1: Total Nitrogen Concentration in Lakes

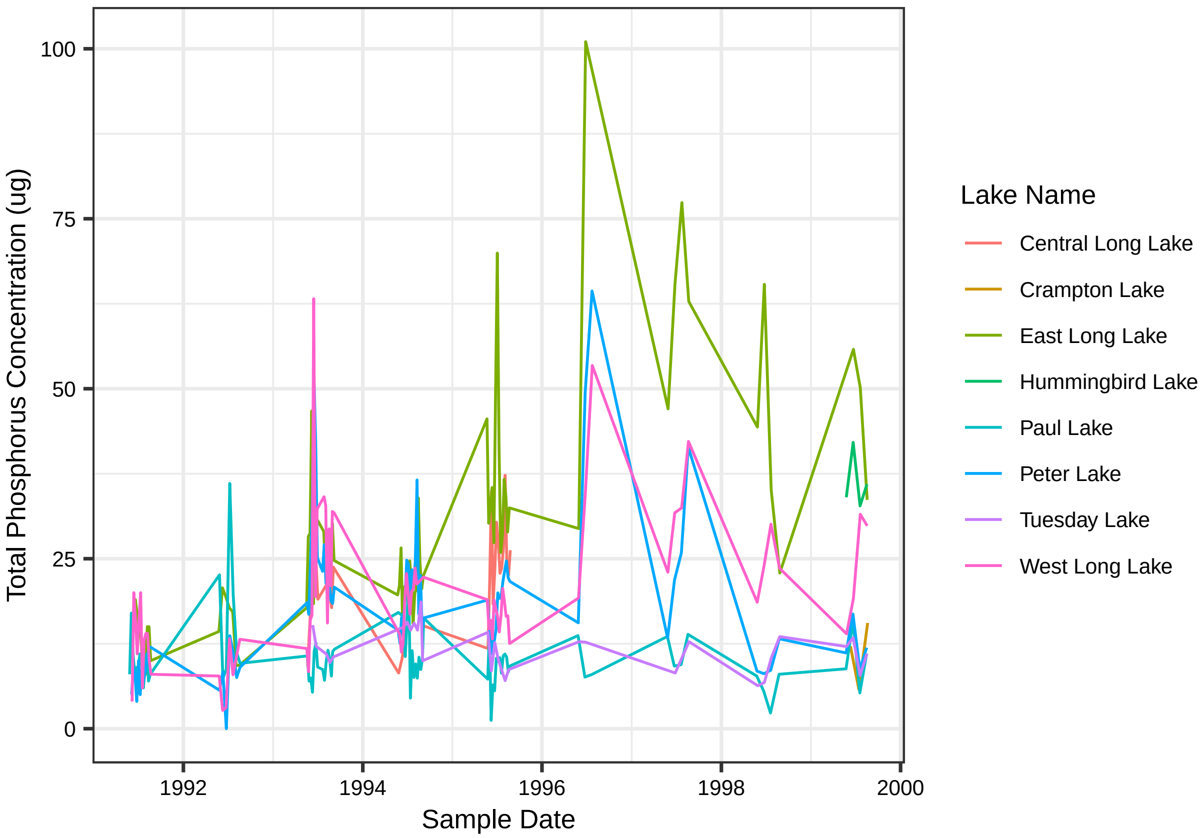


Figure 2: Total Phosphorus Concentration in Lakes

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: