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 creates **organized** and *clean* PDF exports
- It can also allow for just the outputs you want and not run all the code

example * unordered list * item 2 + sub-item 1 + sub-item 2

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	Relevant Information	EDA Use
EPAair_O3_NC2019_raw.csv	air quality monitoring of PM2.5 (pollutant) and ozone in North Carolina by the EPA	Used in Data Wrangling
EPAair_PM25_NC2019_raw.csv	air quality monitoring of PM2.5 and ozone (pollutant) in North Carolina by the EPA	Used in Data Wrangling
NTL-LTER_Lake_Nutrients_Raw.csv	ata from studies on several lakes in the North Temperate Lakes District in Wisconsin, USA by the National Science Foundation	Used in Data Visualization
NTL-LTER_Lake_ChemistryPhysics_Raw.csv	ata from studies on several lakes in the North Temperate Lakes District in Wisconsin, USA by the National Science Foundation	Used in General Linear Modeling

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).

```
#installing a package
install.packages(knitr)
library(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.

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.

```
knitr::opts_chunk$set(message = FALSE, warning = FALSE)
#1. check directory
getwd()
```

```
## [1] "C:/Users/Tasha Griffiths/Documents/Duke Year 1/Spring 22 Classes/Environmental Data Analytics/G:
```

```
#2. load libraries
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(knitr)

#3 Set theme
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

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

```
#filter the dataset
NTL_LTER_Clean <- NTL_LTER %>%
  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>
## <caption>Total Nitrogen Summary</caption>
##   <thead>
##     <tr>
##     <th style="text-align:left;"> lakename </th>
##     <th style="text-align:right;"> Mean_N </th>
##     <th style="text-align:right;"> Min_N </th>
##     <th style="text-align:right;"> Max_N </th>
##     <th style="text-align:right;"> StDev_N </th>
##     </tr>
##   </thead>
## <tbody>
##     <tr>
##     <td style="text-align:left;"> Central Long Lake </td>
##     <td style="text-align:right;"> 690.0469 </td>
##     <td style="text-align:right;"> 343.020 </td>
##     <td style="text-align:right;"> 953.063 </td>
##     <td style="text-align:right;"> 209.09341 </td>
##     </tr>
##     <tr>
```

```
##     <td style="text-align:left;"> Crampton Lake </td>
##     <td style="text-align:right;"> 362.6813 </td>
##     <td style="text-align:right;"> 353.380 </td>
##     <td style="text-align:right;"> 376.304 </td>
##     <td style="text-align:right;"> 12.05748 </td>
##   </tr>
##   <tr>
##     <td style="text-align:left;"> East Long Lake </td>
##     <td style="text-align:right;"> 810.7834 </td>
##     <td style="text-align:right;"> 380.620 </td>
##     <td style="text-align:right;"> 2608.956 </td>
##     <td style="text-align:right;"> 335.41457 </td>
##   </tr>
##   <tr>
##     <td style="text-align:left;"> Hummingbird Lake </td>
##     <td style="text-align:right;"> 1036.6695 </td>
##     <td style="text-align:right;"> 779.053 </td>
##     <td style="text-align:right;"> 1221.960 </td>
##     <td style="text-align:right;"> 204.36889 </td>
##   </tr>
##   <tr>
##     <td style="text-align:left;"> Paul Lake </td>
##     <td style="text-align:right;"> 368.7564 </td>
##     <td style="text-align:right;"> 45.670 </td>
##     <td style="text-align:right;"> 628.625 </td>
##     <td style="text-align:right;"> 106.34741 </td>
##   </tr>
##   <tr>
##     <td style="text-align:left;"> Peter Lake </td>
##     <td style="text-align:right;"> 561.8752 </td>
##     <td style="text-align:right;"> 219.720 </td>
##     <td style="text-align:right;"> 2048.151 </td>
##     <td style="text-align:right;"> 305.64909 </td>
##   </tr>
##   <tr>
##     <td style="text-align:left;"> Tuesday Lake </td>
##     <td style="text-align:right;"> 423.5605 </td>
##     <td style="text-align:right;"> 237.363 </td>
##     <td style="text-align:right;"> 554.418 </td>
##     <td style="text-align:right;"> 78.84522 </td>
##   </tr>
##   <tr>
##     <td style="text-align:left;"> West Long Lake </td>
##     <td style="text-align:right;"> 762.6017 </td>
##     <td style="text-align:right;"> 303.170 </td>
##     <td style="text-align:right;"> 2870.302 </td>
##     <td style="text-align:right;"> 402.95992 </td>
##   </tr>
## </tbody>
## </table>


## <table>
## <caption>Total Phosphorous Summary</caption>
##  <thead>
```

```
##    <tr>
##     <th style="text-align:left;"> lakename </th>
##     <th style="text-align:right;"> Mean_P </th>
##     <th style="text-align:right;"> Min_P </th>
##     <th style="text-align:right;"> Max_P </th>
##     <th style="text-align:right;"> StDev_P </th>
##    </tr>
##   </thead>
## <tbody>
##    <tr>
##     <td style="text-align:left;"> Central Long Lake </td>
##     <td style="text-align:right;"> 21.70981 </td>
##     <td style="text-align:right;"> 8.190 </td>
##     <td style="text-align:right;"> 37.270 </td>
##     <td style="text-align:right;"> 7.076388 </td>
##    </tr>
##    <tr>
##     <td style="text-align:left;"> Crampton Lake </td>
##     <td style="text-align:right;"> 11.16033 </td>
##     <td style="text-align:right;"> 5.803 </td>
##     <td style="text-align:right;"> 15.555 </td>
##     <td style="text-align:right;"> 4.946759 </td>
##    </tr>
##    <tr>
##     <td style="text-align:left;"> East Long Lake </td>
##     <td style="text-align:right;"> 29.28984 </td>
##     <td style="text-align:right;"> 8.000 </td>
##     <td style="text-align:right;"> 101.050 </td>
##     <td style="text-align:right;"> 17.375710 </td>
##    </tr>
##    <tr>
##     <td style="text-align:left;"> Hummingbird Lake </td>
##     <td style="text-align:right;"> 36.21925 </td>
##     <td style="text-align:right;"> 32.765 </td>
##     <td style="text-align:right;"> 42.119 </td>
##     <td style="text-align:right;"> 4.146717 </td>
##    </tr>
##    <tr>
##     <td style="text-align:left;"> Paul Lake </td>
##     <td style="text-align:right;"> 10.45606 </td>
##     <td style="text-align:right;"> 1.222 </td>
##     <td style="text-align:right;"> 36.070 </td>
##     <td style="text-align:right;"> 4.805142 </td>
##    </tr>
##    <tr>
##     <td style="text-align:left;"> Peter Lake </td>
##     <td style="text-align:right;"> 18.39153 </td>
##     <td style="text-align:right;"> 0.000 </td>
##     <td style="text-align:right;"> 64.383 </td>
##     <td style="text-align:right;"> 10.976205 </td>
##    </tr>
##    <tr>
##     <td style="text-align:left;"> Tuesday Lake </td>
##     <td style="text-align:right;"> 11.71853 </td>
```

```
##     <td style="text-align:right;"> 6.325 </td>
##     <td style="text-align:right;"> 18.663 </td>
##     <td style="text-align:right;"> 3.044289 </td>
##   </tr>
##   <tr>
##     <td style="text-align:left;"> West Long Lake </td>
##     <td style="text-align:right;"> 19.82981 </td>
##     <td style="text-align:right;"> 2.690 </td>
##     <td style="text-align:right;"> 63.243 </td>
##     <td style="text-align:right;"> 10.541276 </td>
##   </tr>
## </tbody>
## </table>
```

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.
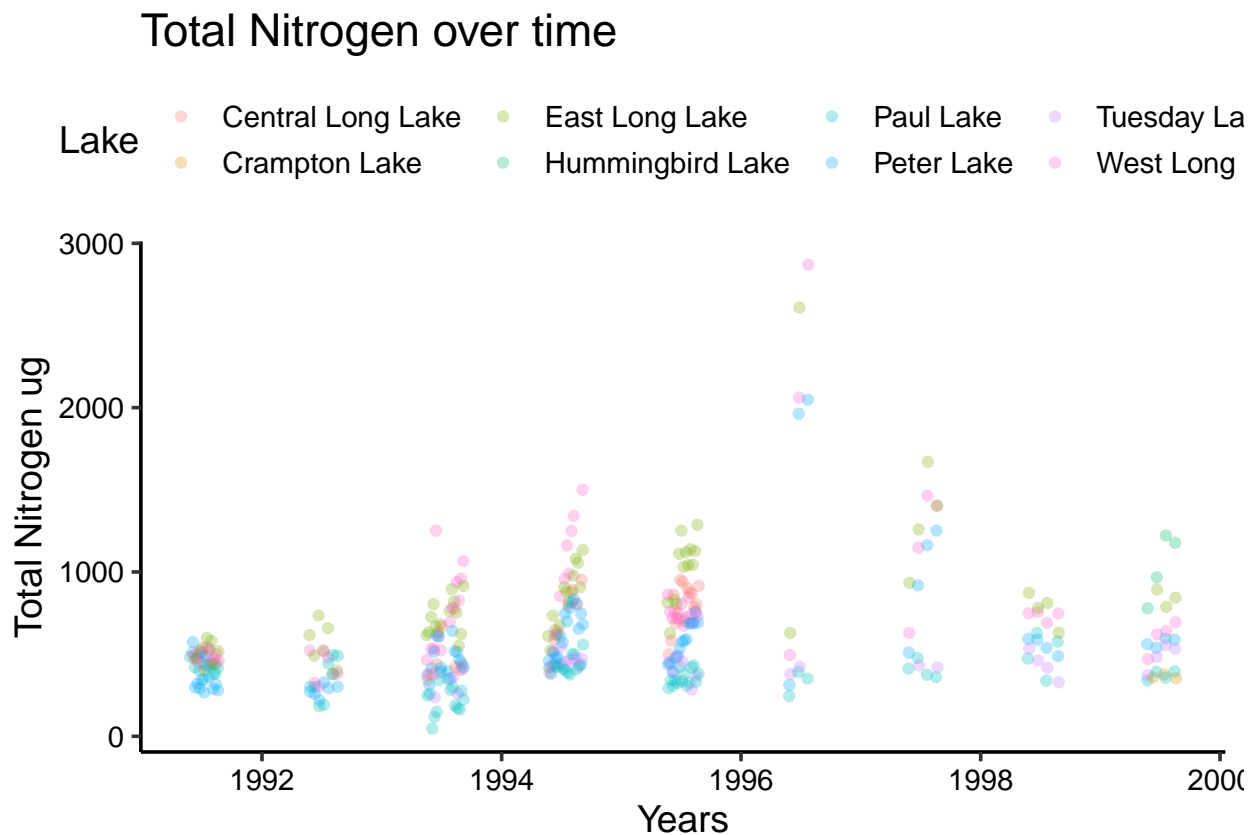
Total Phosphorous over time
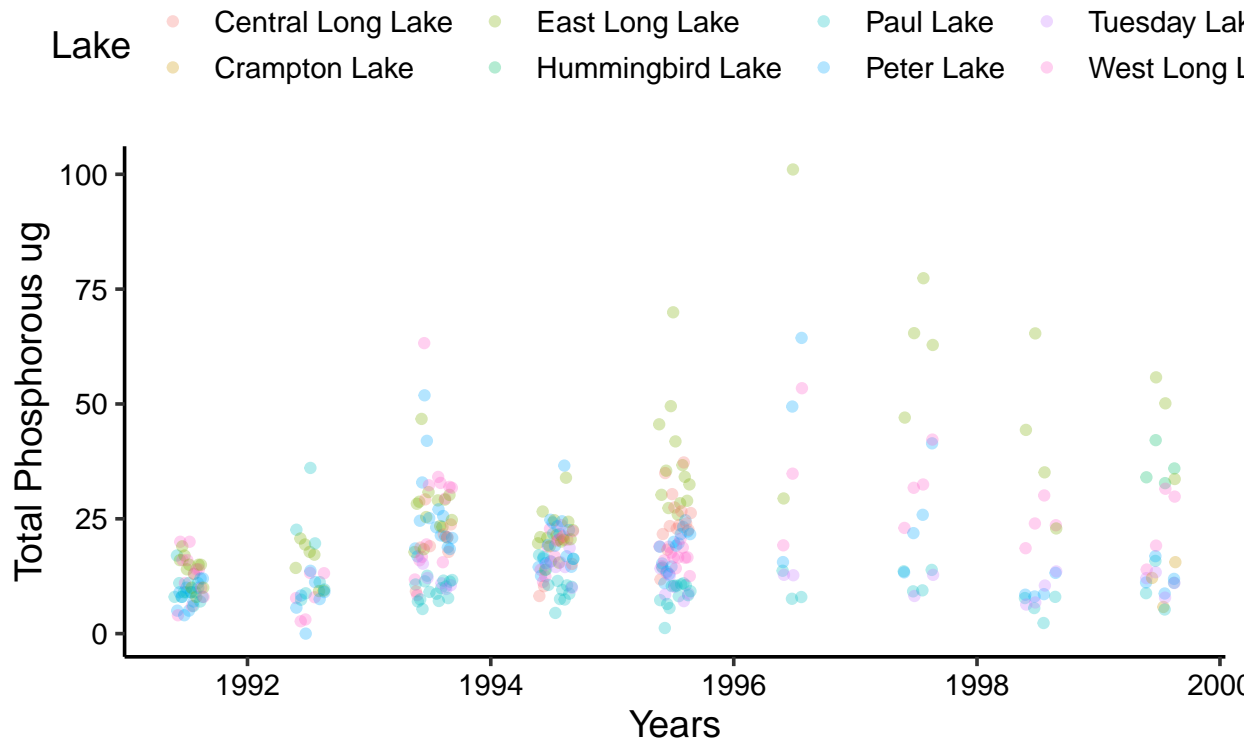
Communicating results

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?

Summary Paragraph

Levels of total Phosphorous and total Nitrogen have changed over time the 8 monitored lakes in Michigan. Phosphorous and Nitrogen are frequently elevated in lake samples due to farmland runoff. However it looks like both chemicals were unver 50 ug in all lakes until about 1996. After 1996 the levels of both chemicals dramatically increased in East Long Lake and Peter and Paul Lakes.

Next Steps for further Analyses

It is interesting that the levels of both Phosphorous and Nitrogen changed dramatically between 1995-1997. In order to understand why - it would be interesting to compare how close each lake is to a neighboring farmland or is close downstream from any chemical processing plants that were built around that time period. Another avenue for analysis would be if any new policies limiting types of fertilizers was removed around that time in the area of these 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: