17: Crafting Reports

Environmental Data Analytics | Kateri Salk

Spring 2019

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

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

```
# Create this R chunk easily by typing: command, option, I
```

3. **Text** with formatting options for readability in knitted document

A handy cheat sheet for R markdown can be found here. Another one can be found here.

WHY R MARKDOWN?

<Fill in our discussion below with bullet points. Use italics and bold for emphasis (hint: use the cheat sheets to figure out how to make bold and italic text).>

- Note: instructions written with <> should not appear in knitted PDF
- *italic*
- **bold**
- An R markdown file allows you to incorporate **figures, tables, and text** all in the same file
- Additionally, this format allows for you to *provide comments* on your code and output so that collaborators can easily understand what you've done and reproduce analyses
- Can easily display the code, output, and accompanying written material in a **PDF format**

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 Dataset Information
EPA Air pollutants measurement	Provides measurements in Ozone and PM

Dataset Name	Relevant Dataset Information
EXOTOX	Provides data from studies on several neonicotinoids and their effects on mortality of various organisms

- Kateri prefers to use the kable function to make tables, this is sometimes glitchy

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` and `knitr`, and sets a ggplot theme. Remember that you need to disable R throwing a message, which containts 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.

```
getwd()
```

```
## [1] "/Users/rachelgonsenhauser/Documents/Environmental_Data_Analytics_2020"
```

```
library(tidyverse)
library(knitr)

mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)

Nutrients.raw <- read.csv("./Data/Raw/NTL-LTER_Lake_Nutrients_Raw.csv")
head(Nutrients.raw)
```

```
##   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
```

```
Nutrients.raw$sampledate <- as.Date(Nutrients.raw$sampledate, format = "%m/%d/%y")
class(Nutrients.raw$sampledate)
```

```
## [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)

```
Nutrients.processed <-
  Nutrients.raw %>%
  select(lakename:sampledate, depth:po4) %>%
  filter(depth == 0)

#could also code like this to exlucde rows we don't want
#Nutrients.processed <-
  #Nutrients.raw %>%
  #select(-lakeid,-depth_id,-comments) %>%
  #filter(depth == 0)
```

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.

```
Nitrogen.summary <-
  Nutrients.processed %>%
  drop_na() %>%
  group_by(lakename) %>%
  summarize(meanTN = mean(tn_ug),
            minTN = min(tn_ug),
            maxTN = max(tn_ug),
            sdTN = sd(tn_ug))

Phosphorous.summary <-
  Nutrients.processed %>%
  drop_na() %>%
  group_by(lakename) %>%
  summarize(meanTP = mean(tp_ug),
            minTP = min(tp_ug),
            maxTP = max(tp_ug),
            sdTP = sd(tp_ug))
```

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.

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.

```
library(scales)
```

```
##
```

```
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
ggplot(Nutrients.processed, aes(x = as.POSIXct(sampledate), y = tn_ug, color = lakename)) +
  geom_point() +
  scale_color_brewer(palette="Dark2") +
  xlab("Date (year)") +
  ylab("Total Nitrogen (ug/L)") +
  scale_x_datetime(date_breaks = "1 year", labels = date_format("%Y"))
```

```
## Warning: Removed 139 rows containing missing values (geom_point).
```

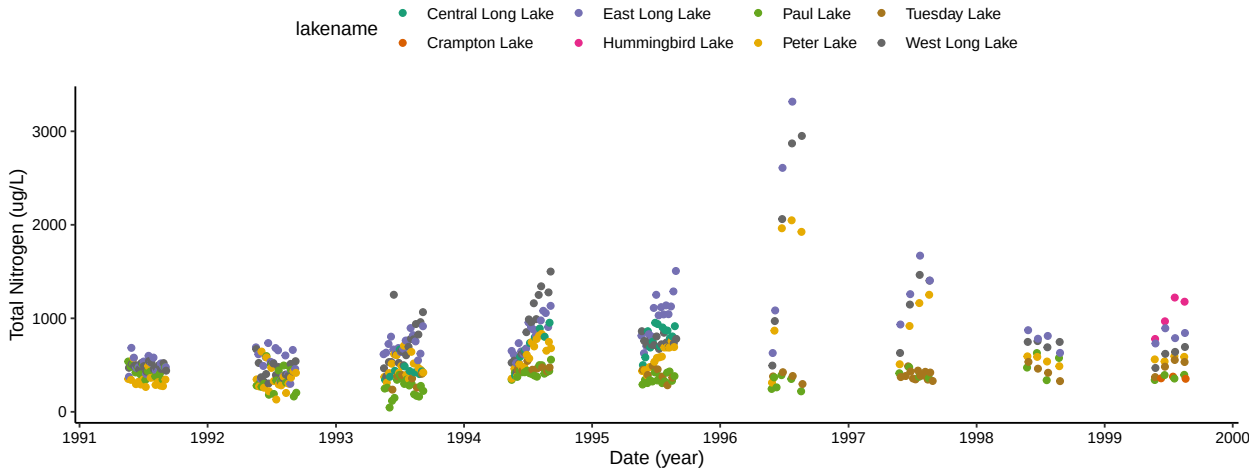
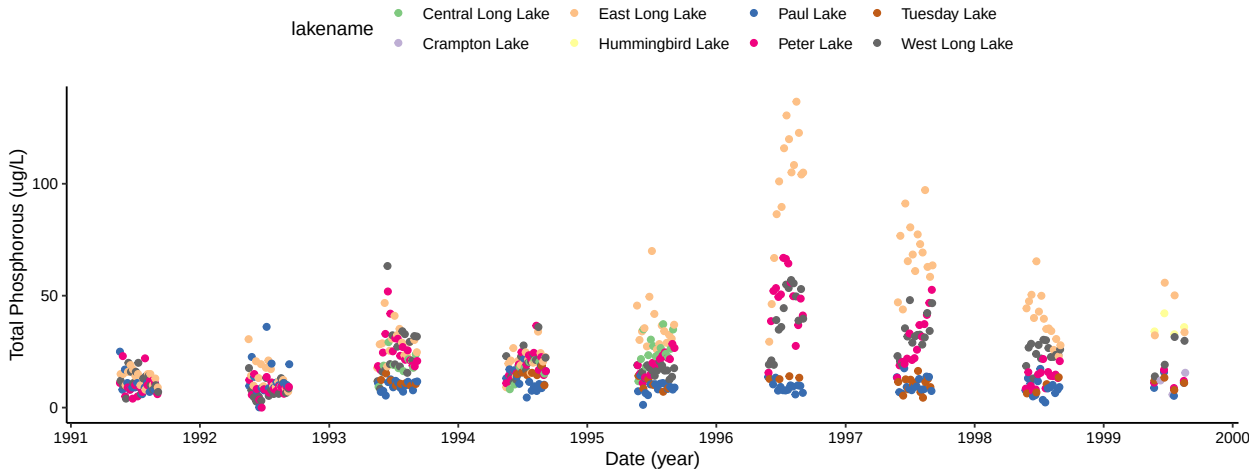


Figure 1: Total nitrogen over time

```
library(scales)
ggplot(Nutrients.processed, aes(x = as.POSIXct(sampledate), y = tp_ug, color = lakename)) +
  geom_point() +
  scale_color_brewer(palette="Accent") +
  xlab("Date (year)") +
  ylab("Total Phosphorous (ug/L)") +
  scale_x_datetime(date_breaks = "1 year", labels = date_format("%Y"))
```

```
## Warning: Removed 7 rows containing missing values (geom_point).
```



Other options What are the chunk options that will suppress the display of errors, warnings, and messages in the final document?

> ANSWER: * suppress display of warnings - warnings=FALSE * suppress display of messages - messages=FALSE

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?

Total phosphorous concentrations experience a peak between 1996 and 1997, where East Long Lake has the highest concentrations among lakes compared. Total nitrogen also peaks around this same time, with East and West Long Lakes experiencing the highest concentrations.

Next steps for research might include gathering more data for other contaminants included in the dataset (such as NH3^4, NO2^3, and PO4). These variables have a lot of missing data, so gathering more data would allow us to compare concentrations of total nitrogen and phosphorous to the occurrence of these other chemical constituents. To analyze these occurrences, a multiple linear regression could be used to see if total nitrogen or phosphorous are predictors for the occurrence of other compounds, or vice versa.

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: