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
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?

- Code, output and test/notes together in one document
- Knit to useful formats(pdf, html, docx)
- Legible code + output
- Git friendly
- Reproducible
- Flexible formatting
- Simple syntax and autoreferencing

TEXT EDITING CHALLENGE

Create a table below that details the example datasets we have been using in class. The first column should contain the name of the dataset and the second column should include some relevant information about the dataset.

DatasetName	Infomation
EPAair	Air quality
NTL-LTER	Lake nutrients

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` and `knitr`, and sets a ggplot 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.

```
getwd()
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------- tidyve
```

```
## v ggplot2 3.1.0       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

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

```
Lake <- read.csv("./Data/Raw/NTL-LTER_Lake_Nutrients_Raw.csv")
head(Lake)
Lake$sampledate <- as.Date(Lake$sampledate, format = "%m/%d/%y")
```

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)

```
colnames(Lake)
```

```
##  [1] "lakeid"     "lakename"   "year4"      "daynum"     "sampledate"
##  [6] "depth_id"   "depth"      "tn_ug"      "tp_ug"      "nh34"
## [11] "no23"       "po4"        "comments"
```

```
Lake_tidy <- Lake %>%
  select("lakename","year4","daynum","sampledate","depth","tn_ug","tp_ug", "nh34","no23","po4")%>%
  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.

```
Lake_summary_tn <- subset(Lake, !is.na(tn_ug)) %>%
  group_by(lakename)%>%
  summarise(mean_tn = mean(tn_ug), min_tn = min(tn_ug), max_tn = max(tn_ug))
Lake_summary_tp <- subset(Lake, !is.na(tn_ug)) %>%
  group_by(lakename)%>%
  summarise(mean_tp = mean(tp_ug), min_tp = min(tp_ug), max_tp = max(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.

Table 2: Total N

lakename	mean_tn	min_tn	max_tn
Bergner Lake	471.3840	360.5784	626.5504
Bolger Bog	800.5791	647.7846	1334.3991
Brown Lake	667.4650	390.8921	1094.6642
Central Long Lake	794.4133	157.1900	2474.3030
Crampton Lake	351.9243	163.3900	956.4060
Cranberry Bog	414.4075	355.2214	494.5169
East Long Lake	848.9101	0.0000	3316.8920
Hummingbird Lake	915.1903	612.6930	1462.5070
Inkpot Lake	464.0169	390.2457	549.1784
Morris Lake	639.8115	545.4971	767.4801
North Gate Bog	498.4990	412.3507	589.2487
Paul Lake	433.3314	45.6700	2099.0000
Peter Lake	534.3640	111.2500	3497.6990
Plum Lake	392.4660	324.6816	447.4974
Raspberry Lake	394.4905	368.8612	426.0130
Reddington Lake	668.8188	583.0434	790.9104
Roach Lake	253.6822	229.4159	287.1464
Tender Bog	545.2030	504.5756	587.6459
Tenderfoot Lake	461.6497	359.4719	615.7022
Tuesday Lake	532.9443	215.4970	1572.2620
Ward Lake	488.7789	365.1683	658.2269
West Long Lake	753.3605	155.6100	2950.3430

Table 3: Total P

lakename	mean_tp	min_tp	max_tp
Bergner Lake	11.358813	8.311631	14.45676
Bolger Bog	58.538284	32.512804	154.77538
Brown Lake	34.939876	22.629511	55.72426
Central Long Lake	NA	NA	NA
Crampton Lake	NA	NA	NA
Cranberry Bog	12.374829	8.603880	15.34540
East Long Lake	NA	NA	NA
Hummingbird Lake	34.358315	17.810702	68.40200
Inkpot Lake	19.173572	11.895689	27.37257
Morris Lake	23.580105	18.306052	31.46582
North Gate Bog	14.963208	10.842851	19.36674

lakename	mean_tp	min_tp	max_tp
Paul Lake	NA	NA	NA
Peter Lake	NA	NA	NA
Plum Lake	16.473345	8.777133	22.15730
Raspberry Lake	12.702892	9.804985	16.07796
Reddington Lake	17.759235	13.974725	23.64643
Roach Lake	9.702002	4.808682	17.82526
Tender Bog	12.722948	9.792445	16.67473
Tenderfoot Lake	19.651588	5.151224	40.11878
Tuesday Lake	NA	NA	NA
Ward Lake	24.046563	18.939480	35.12000
West Long Lake	NA	NA	NA

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.

Other options

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

ANSWER:

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?

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: