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?

- please
- put
- space!
- Code, output, and text/notes together in one document
- Knit to useful formats (pdf, html, docx)
- Legible code and output
- Git friendly - version control
- Reproducible
- Updating capabilities
- Focus on output and conclusions, not code (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.

Dataset	Source

Dataset	Source
NTL-LTER_Lake	North Temperate Lakes Long Term Ecological Research
EPAair_PM25_NC2018	U.S. Environmental Protection Agency

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.

```
## [1] "/Users/Tristen/OneDrive - Duke University/Spring 2019/Data Analytics/Environmental_Data_Analyti
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

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

Load the NTL-LTER_Lake_Nutrients_Raw dataset, display the head of the dataset, and set the date column to a date format.

```
##   lakeid  lakename year4 daynum sampledate depth_id depth tn_ug tp_ug nh34
## 1      L Paul Lake  1991    140    5/20/91        1  0.00   538    25   NA
## 2      L Paul Lake  1991    140    5/20/91        2  0.85   285    14   NA
## 3      L Paul Lake  1991    140    5/20/91        3  1.75   399    14   NA
## 4      L Paul Lake  1991    140    5/20/91        4  3.00   453    14   NA
## 5      L Paul Lake  1991    140    5/20/91        5  4.00   363    13   NA
## 6      L Paul Lake  1991    140    5/20/91        6  6.00   583    37   NA
##   no23 po4 comments
## 1   NA  NA
## 2   NA  NA
## 3   NA  NA
## 4   NA  NA
## 5   NA  NA
## 6   NA  NA
```

Customize the chunk options such that the code is run but is not displayed in the final document.

Table 2: Total Nitrogen Summary

lakename	meanTN	maxTN	minTN	sdTN
Bergner Lake	471.3840	626.5504	360.5784	92.52036
Bolger Bog	800.5791	1334.3991	647.7846	197.59391
Brown Lake	667.4650	1094.6642	390.8921	185.81284
Central Long Lake	794.4133	2474.3030	157.1900	510.04678
Crampton Lake	351.9243	956.4060	163.3900	137.38049
Cranberry Bog	414.4075	494.5169	355.2214	47.42169
East Long Lake	848.9101	3316.8920	0.0000	492.11923
Hummingbird Lake	915.1903	1462.5070	612.6930	200.34164
Inkpot Lake	464.0169	549.1784	390.2457	57.29937
Morris Lake	639.8115	767.4801	545.4971	80.28057
North Gate Bog	498.4990	589.2487	412.3507	50.09471
Paul Lake	433.3314	2099.0000	45.6700	308.23787
Peter Lake	534.3640	3497.6990	111.2500	400.92843
Plum Lake	392.4660	447.4974	324.6816	45.37608
Raspberry Lake	394.4905	426.0130	368.8612	20.33686
Reddington Lake	668.8188	790.9104	583.0434	67.51347
Roach Lake	253.6822	287.1464	229.4159	17.08657
Tender Bog	545.2030	587.6459	504.5756	42.13848
Tenderfoot Lake	461.6497	615.7022	359.4719	80.55970
Tuesday Lake	532.9443	1572.2620	215.4970	211.69369
Ward Lake	488.7789	658.2269	365.1683	73.22381
West Long Lake	753.3605	2950.3430	155.6100	489.35476

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)

```
NTL_Processed <- NTL_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.

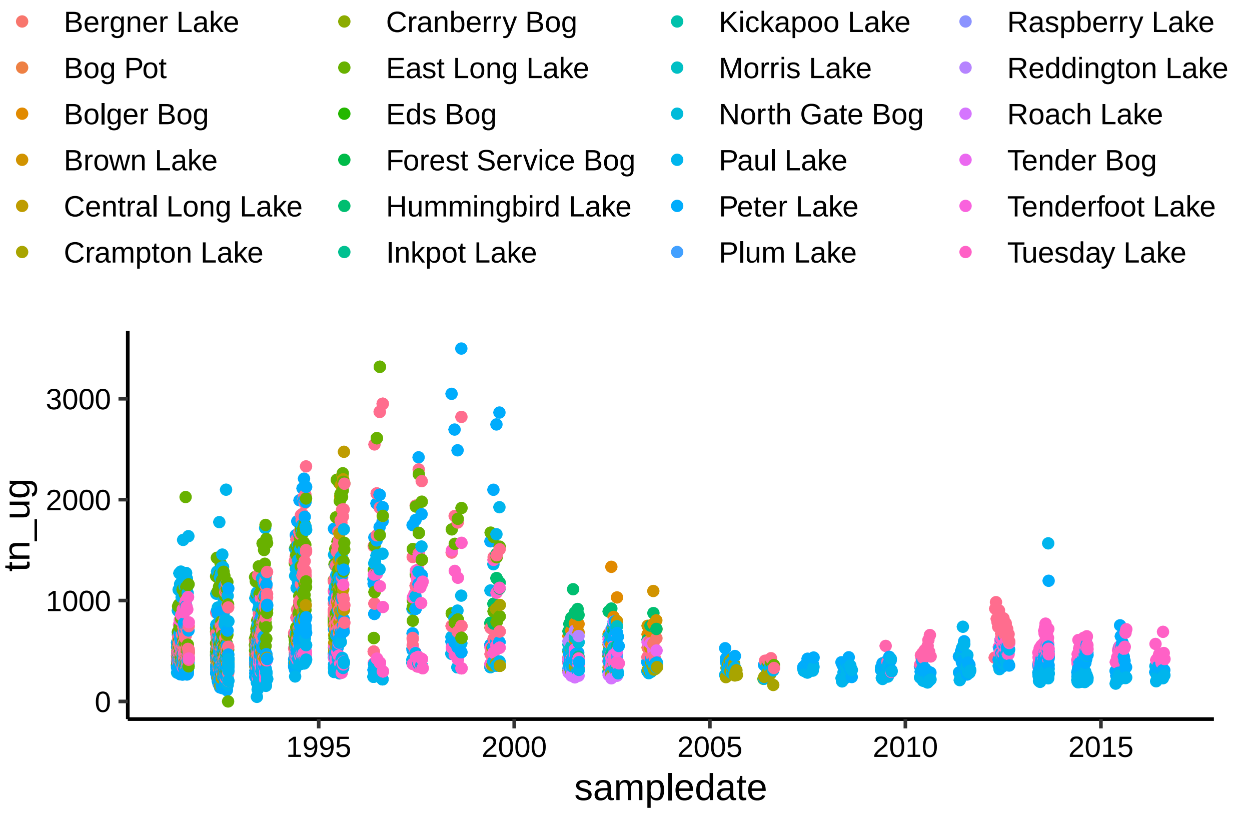
Create a third R chunk that uses the function `kable` in the knitr package to display two tables: one for the summary dataframe for total N and one for the summary dataframe of total P. Use the `caption = " "` code within that function to title your tables. Customize the chunk options such that the final table is displayed but not the code used to generate the table.

```
kable(NTL.Summary.Nitrogen, "latex", caption = "Total Nitrogen Summary", booktabs = T)
```

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.

```
ggplot(NTL_Raw, aes(x = sampledate, y = tn_ug, color = lakename)) +
  geom_point()
```

```
## Warning: Removed 2330 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:

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:
```