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‘ (‘control’, ‘option’, ‘I’ on Macs)
    - 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 that can be commented
- Git friendly - version control!
- Reproducible
- Updating capabilities
- Focus on output and conclusions, not code (or vice versa) (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	Description
Ecotox	Mortality of organisms from Neonicotinoids
EPA Air	O3 and PM2.5 levels from 2017 and 2018
NTL-LTER	Nutrient and Chemical data for lakes in Wisconsin
USGS	Flow rates from Eno River Spring 2019

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.

```
## [1] "/Users/lindsayroth/Documents/MEM 1st Year/Spring 2019/Env_Data_Analytics/Env_Data_Analytics"
```

```
## -- Attaching packages ---------------------------------------------------------------- tidyverse 1.2.
```

```
## 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_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
##   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
```

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 %>%
  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.

```
Nutrients_Summary_N <- Nutrients %>%
  filter(!is.na(tn_ug)) %>%
  group_by(lakename) %>%
  summarize(tn_mean = mean(tn_ug), tn_min = min(tn_ug),
            tn_max = max(tn_ug), tn_std = sd(tn_ug))

Nutrients_Summary_P <- Nutrients %>%
  filter(!is.na(tp_ug))  %>%
  group_by(lakename) %>%
  summarize(tp_mean = mean(tp_ug), tp_min = min(tp_ug),
            tp_max = max(tp_ug), tp_std = 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.

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

lakename	tp_mean	tp_min	tp_max	tp_std
Bergner Lake	12.247609	8.3116307	23.54200	3.619843
Bog Pot	42.700000	29.2370000	59.09800	13.564224
Bolger Bog	52.317400	26.2790000	154.77538	32.941216
Brown Lake	41.247157	22.6295107	80.70900	17.737526
Central Long Lake	27.650641	3.8700000	86.76200	15.055482
Crampton Lake	11.519615	3.2700000	53.33300	8.873020
Cranberry Bog	13.793455	8.3289575	28.54400	5.101257
East Long Lake	58.782476	0.0000000	352.05600	62.109048
Eds Bog	36.614500	33.7970000	40.75000	3.293310
Forest Service Bog	17.532250	13.2480000	23.15600	4.169864
Hummingbird Lake	34.376353	17.8107015	69.86900	10.788058
Inkpot Lake	21.656568	11.8956886	37.88900	7.433935
Kickapoo Lake	32.879250	25.9960000	37.20000	4.857109
Morris Lake	24.331630	18.3060521	34.51800	4.835955
North Gate Bog	15.723024	10.1473971	30.81200	5.288118
Paul Lake	16.097701	-6.3488500	123.29000	16.711673
Peter Lake	25.209163	-3.0970000	157.25000	23.316691
Plum Lake	18.430711	8.4414848	35.98800	7.037070
Raspberry Lake	13.220990	9.8049848	20.48700	2.965015
Reddington Lake	18.677777	11.1614607	35.10200	6.761786
Roach Lake	9.948343	4.8086825	17.82526	3.652945
Tender Bog	12.722948	9.7924446	16.67473	3.290124
Tenderfoot Lake	28.294714	5.1512240	90.21764	20.363216
Tuesday Lake	20.310104	-0.7783225	101.89000	12.926009
Ward Lake	25.048789	18.9394796	35.12000	4.610931
West Long Lake	37.010093	2.6900000	251.67400	39.090908

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: