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 FRA: yust another ar markdown language
    - 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?

- Create a bullet
- Other way of ceatitng a bullet
- Other way but tricky becuase if you dont put a space
- Code, output, and test/notes together in one document
- Knit to useful formats (pdf, html, docx)
- Legible code + 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	Information
ECOTOX Neonicotinoid	Contains data from studies on several neonicotinoids and their effects on mortality of various organisms.
EPA Air Quality	Contains data from air quality monitoring of PM2.5 and ozone in North Carolina in 2017 and 2018.

Dataset	Information
NTL-LTER Lake	Contains data from studies on several lakes in the North Temperate Lakes District in Wisconsin, USA. Data were collected as part of the Long Term Ecological Research station established by the National Science Foundation.
USGS Streamflow data for site 02085000	Contains streamflow data from the USGS streamflow gage site 02085000 (Eno River at Hillsborough, NC).

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.

```
getwd()
```

```
## [1] "C:/Users/Felipe/OneDrive - Duke University/1. DUKE/1. Ramos 2 Semestre/EOS-872 Env. Data Analyt
```

```
library(tidyverse)
```

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

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

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

```
library(knitr)

felipe_theme <- theme_light(base_size = 12) +
  theme(axis.text = element_text(color = "grey8"),
        legend.position = "right", plot.title = element_text(hjust = 0.5))
theme_set(felipe_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.

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)

```r
NTL_LTER_Lake_Nutrients_Processed <-
  NTL_LTER_Lake_Nutrients_Raw %>%
  select(lakename, year4, daynum, sampledate, depth, tn_ug, tp_ug, nh34, no23, po4) %>%
  filter(depth == 0) %>%
  filter(!is.na(tn_ug)) %>%
  filter(!is.na(tp_ug))
```

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 2: Mean, minimum, maximum, and standard deviation of total nitrogen concentrations for each lake

lakename	mean_tn_ug	min_tn_ug	max_tn_ug	sd_tn_ug
Central Long Lake	675.8338	343.020	953.063	203.25838
Crampton Lake	362.6813	353.380	376.304	12.05748
East Long Lake	794.3737	299.310	3316.892	414.98782
Hummingbird Lake	1036.6695	779.053	1221.960	204.36889
Paul Lake	365.0360	45.670	628.625	107.86320
Peter Lake	548.2733	131.830	2048.151	320.83105
Tuesday Lake	410.0794	237.363	554.418	72.71582
West Long Lake	737.8763	303.170	2950.343	438.44999

Table 3: Mean, minimum, maximum, and standard deviation of total phosphorus concentrations for each lake

lakename	mean_tp_ug	min_tp_ug	max_tp_ug	sd_tp_ug
Central Long Lake	21.16577	8.190	37.270	6.747806
Crampton Lake	11.16033	5.803	15.555	4.946759
East Long Lake	27.98533	7.160	119.932	19.137657
Hummingbird Lake	36.21925	32.765	42.119	4.146717
Paul Lake	10.59191	0.110	36.070	4.854132
Peter Lake	17.79234	0.000	64.383	10.965644
Tuesday Lake	11.37014	4.413	18.663	3.141466
West Long Lake	18.45639	2.690	63.243	10.488876

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:
```