



MAY 18-29

Data science as an entryway to open publishing

Julia Stewart Lowndes & Nicholas Tierney

[@juliesquid](https://twitter.com/juliesquid) & [@nj_tierney](https://twitter.com/nj_tierney)

2020-05-27

CC BY 4.0

openpublishingfest.org

slides: openscapes.org/media

Hi, we're Julie & Nick



Julia Stewart Lowndes, PhD
[@juliesquid](https://twitter.com/juliesquid)

Marine data scientist, Openscapes lead
National Center for Ecological Analysis
& Synthesis, UC Santa Barbara, USA



Nicholas Tierney, PhD
[@nj_tierney](https://twitter.com/nj_tierney)

Lecturer in Statistics
Monash University, Australia

On Twitter and IRL, we
are active members of
the [#rstats](https://twitter.com/rstats) community:

- [@rOpenSci](https://twitter.com/rOpenSci)
- [@RLadies](https://twitter.com/RLadies)
- [@RStudio](https://twitter.com/RStudio)

I came to R for the
data analysis,
and was blown away
by the publishing



**The same workflow you use for data analysis
– rooted in reproducibility – empowers you
make your work available to the world**

...in ways you never imagined

Using RMarkdown for scientific publishing

Fueling reproducibility in data science

RMarkdown

RMarkdown powerfully combines executable R code with simple text formatting and for efficient, automatable, reproducible research

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
```${r cars}
summary(cars)
```
```

Including Plots

You can also embed plots, for example:

```
```${r pressure, echo=FALSE}
plot(pressure)
```
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Simple text formatting

R code

Analyses and figures are in the same place as your reporting document:
saves time as you iterate!

Enables good practices for reproducibility & versioning

```
---
title: "Untitled"
author: "Julie Lowndes"
date: "5/26/2020"
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the Knit button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

## Including Plots

You can also embed plots, for example:

```{r pressure, echo=FALSE}
plot(pressure)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
```

Our RMarkdown file renders to:

Word!

PDF!

Untitled

Julie Lowndes
5/26/2020

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

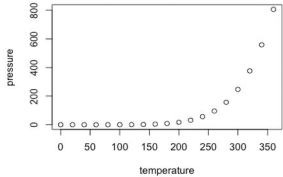
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

| ## | speed | dist |
|-------------|-------|----------------|
| ## Min. : | 4.0 | Min. : 2.00 |
| ## 1st Qu.: | 12.0 | 1st Qu.: 26.00 |
| ## Median : | 15.0 | Median : 36.00 |
| ## Mean : | 15.4 | Mean : 42.98 |
| ## 3rd Qu.: | 19.0 | 3rd Qu.: 56.00 |
| ## Max. : | 25.0 | Max. : 120.00 |

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Untitled

Julie Lowndes
5/26/2020

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

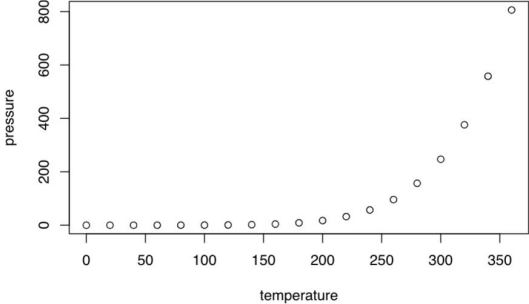
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

| ## | speed | dist |
|-------------|-------|----------------|
| ## Min. : | 4.0 | Min. : 2.00 |
| ## 1st Qu.: | 12.0 | 1st Qu.: 26.00 |
| ## Median : | 15.0 | Median : 36.00 |
| ## Mean : | 15.4 | Mean : 42.98 |
| ## 3rd Qu.: | 19.0 | 3rd Qu.: 56.00 |
| ## Max. : | 25.0 | Max. : 120.00 |

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Imagine never copy-pasting a graph into your report again!!!!

RMarkdown can also manage citations, cross-referencing figures and section headers.

Using RMarkdown beyond your wildest dreams

Reimagining sharing and publishing online

RMarkdown

RMarkdown creates HTML files that can be shared openly on the web

```
---
title: "Untitled"
author: "Julie Lowndes"
date: "5/26/2020"
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the Knit button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

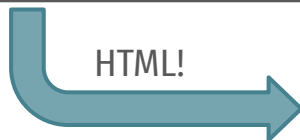
## Including Plots

You can also embed plots, for example:

```{r pressure, echo=FALSE}
plot(pressure)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
```

Our RMarkdown file renders to:



Untitled

Julie Lowndes
5/26/2020

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed      dist
## Min.   : 4.0   Min.   : 2.00
## 1st Qu.:112.0 1st Qu.: 26.00
## Median :15.0  Median : 36.00
## Mean   :15.4   Mean   : 42.98
## 3rd Qu.:119.0 3rd Qu.: 56.00
## Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:

```
##      temperature      pressure
## Min.   : 0.000000000  Min.   : 0.000000000
## 1st Qu.: 50.000000000 1st Qu.: 0.000000000
## Median :100.000000000 Median : 0.000000000
## Mean   :150.000000000 Mean   : 0.000000000
## 3rd Qu.:200.000000000 3rd Qu.: 0.000000000
## Max.   :350.000000000 Max.   : 800.000000000
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

We can store and distribute html files on **GitHub**, which also offers display options for publishing.

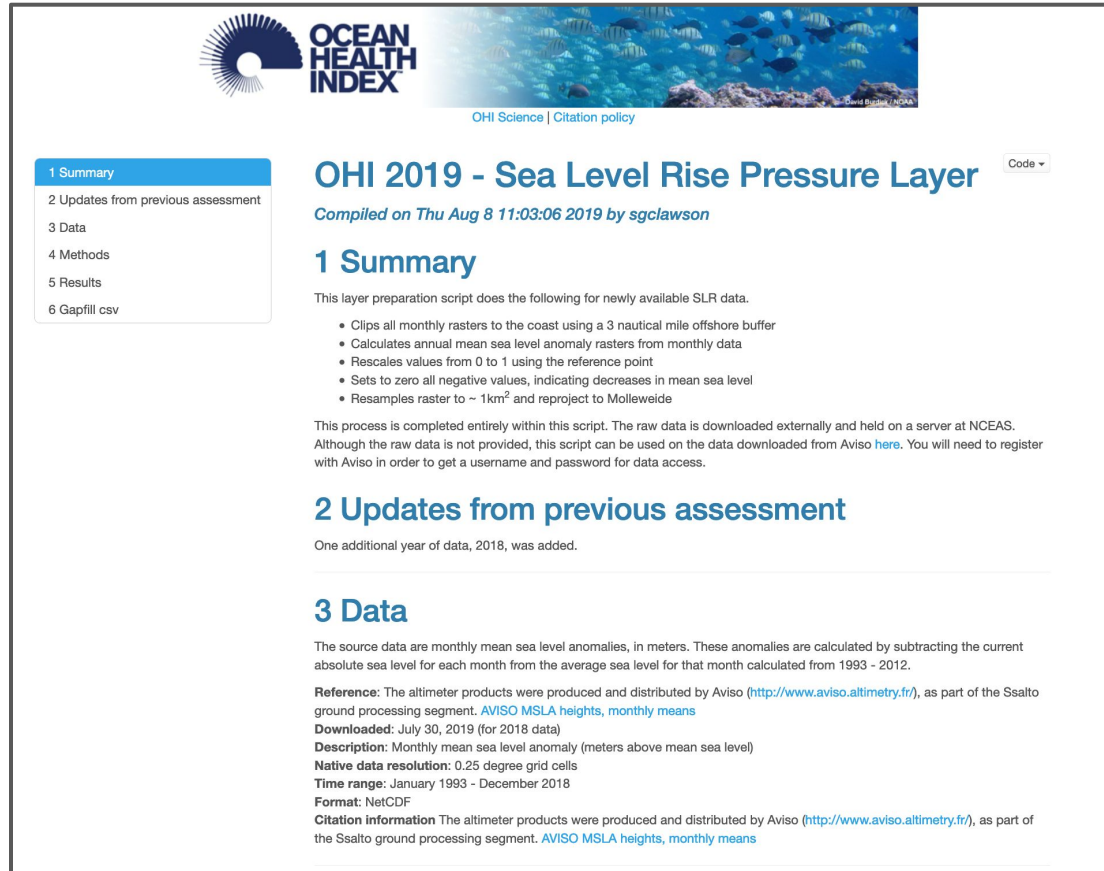
Let's look at some real-world examples from science...

Suddenly you can share a URL rather than attaching a file!

And that same URL will update rather than re-attaching a new version of the file!

Single-page html

RMarkdown html files for open publishing; URL will display most recent version



OCEAN HEALTH INDEX

OHI Science | Citation policy

OHI 2019 - Sea Level Rise Pressure Layer

Code ▾

Compiled on Thu Aug 8 11:03:06 2019 by sgclawson

1 Summary

This layer preparation script does the following for newly available SLR data.

- Clips all monthly rasters to the coast using a 3 nautical mile offshore buffer
- Calculates annual mean sea level anomaly rasters from monthly data
- Rescales values from 0 to 1 using the reference point
- Sets to zero all negative values, indicating decreases in mean sea level
- Resamples raster to ~ 1km² and reproject to Molleweide

This process is completed entirely within this script. The raw data is downloaded externally and held on a server at NCEAS. Although the raw data is not provided, this script can be used on the data downloaded from Aviso [here](#). You will need to register with Aviso in order to get a username and password for data access.

2 Updates from previous assessment

One additional year of data, 2018, was added.

3 Data

The source data are monthly mean sea level anomalies, in meters. These anomalies are calculated by subtracting the current absolute sea level for each month from the average sea level for that month calculated from 1993 - 2012.

Reference: The altimeter products were produced and distributed by Aviso (<http://www.aviso.altimetry.fr/>), as part of the Ssalto ground processing segment. [AVISO MSLA heights, monthly means](#)

Downloaded: July 30, 2019 (for 2018 data)

Description: Monthly mean sea level anomaly (meters above mean sea level)

Native data resolution: 0.25 degree grid cells

Time range: January 1993 - December 2018

Format: NetCDF

Citation information The altimeter products were produced and distributed by Aviso (<http://www.aviso.altimetry.fr/>), as part of the Ssalto ground processing segment. [AVISO MSLA heights, monthly means](#)

Examples from the [Ocean Health Index](#)

Many display options;
floating table of contents,
show/hide code

Then we can think about
organization & discoverability: How to
organize multiple htmls?
And how do we find them?

ohi-science.org/ohiprep_v2019/globalprep/prs_slr/v2019/slr_layer_prep_v2.html

Learn: rmarkdown.rstudio.com

Simple Websites

Combine RMarkdown files as a website with a navigation bar between pages, requires only GitHub

GLOBAL

Goals

Layers

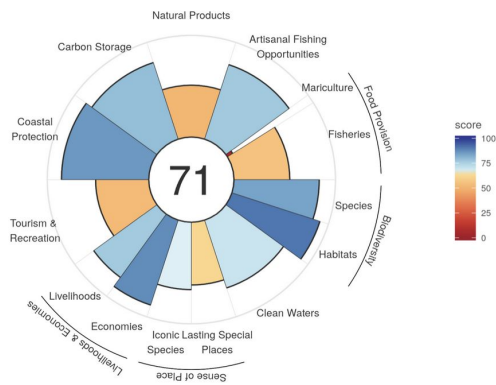
Scores

Download

Fellows

Annual OHI global assessments

Ocean Health Index scores provide invaluable, comprehensive, and quantitative assessments of progress towards healthy and sustainable oceans. Such assessments are particularly valuable when repeated annually. 2019 marks the eighth year of annual global Ocean Health Index (OHI) assessments, with scores representing ocean health for 220 coastal nations and territories. For detailed description of our up-to-date data and methods, see the [Supplemental Methods](#). And visit our [Story Map](#)!



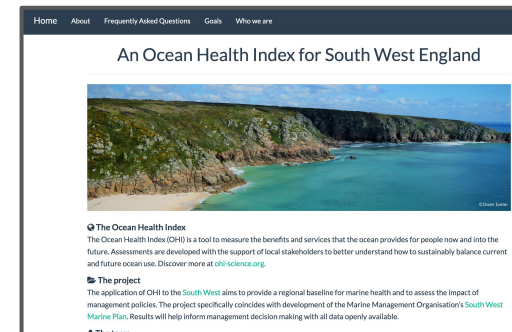
The average 2019 Index score was 71 out of 100. Average Index scores have not dramatically changed over eight years, which could be expected at a global scale. However, some individual goals and regions have had significant changes.

- [Learn about the goals](#)
- [Read the 2019 results summary](#)
- [Explore scores by region and goal](#)
- [Download data](#)
- [Dig deeper with the Supplemental Methods](#)
- [Learn more at ohi-science.org](#)

The OHI framework has not changed since its inception in 2012 (*Halpern et al. 2012*). We continue to assess ocean health based on the sustainable delivery of a suite of goals that are important to humans. We score each goal on a scale of 0-100 based on current

Useful for organizing, e.g. linking out to additional single-page htmls

You can also create templates and populate them automatically



ohi-science.org/ohi-global


ohi-science.org/esw

Learn: jules32.github.io/rmarkdown-website-tutorial

Blogdown Websites

Create powerful websites with more complexity and blogging capabilities; requires more setup & deployment from a server

BLOG TALKS PROJECTS ABOUT / CONTACT 🔍 🌙



Alison Hill
Data Scientist & Professional Educator
RStudio

[🔗](#) [🐦](#) [👤](#) [📄](#) [🆔](#) [in](#) [🐙](#)
[🔍](#)

I am a Data Scientist & Professional Educator at **RStudio PBC**. I am an **international keynote speaker**, and I regularly lead workshops and develop online learning materials on topics like **reproducible research**, **machine learning**, and data visualization. My teaching materials have been used by **NASA**, **Pew Research Center**, **University of Oregon**, and now **RStudio**. I am also a co-author of the book *blogdown: Creating Websites with R Markdown*.

I received my PhD in psychology and quantitative methods from Vanderbilt University in 2008. Prior to joining RStudio, I was an Assistant Professor at Arizona State University, and an Associate Professor at Oregon Health & Science University (OHSU). While at OHSU, I was an NIH-funded Principal Investigator and the Assistant Director of the *Center for Spoken Language Understanding*. I was nominated for a distinguished faculty award for outstanding teaching, and was awarded an **excellence in graduate education award** from the OHSU School of Medicine. My research has been published in **Pediatrics**, **Autism Research**, and other **peer-reviewed journals**.

Interests

- Data science
- Statistics
- Predictive modeling & machine learning
- Reproducible research
- Education
- Parenting

Education

- 🎓 PhD in Developmental Psychology & Quantitative Methods, 2008
Vanderbilt University
- 🎓 MSc in Developmental Psychology, 2005
Vanderbilt University
- 🎓 BSc in Applied Psychology, 2002
Georgia Institute of Technology

“If you want to learn to write, you read a lot, if you want to play music, you listen a lot. It’s hard to do this with data analysis.” - [Hilary Parker & Roger Peng, RStudio::conf\(2020\) keynote](#)

So we write blogs and tutorials to share code, discuss, and learn together.

Power to organize, tag, search, navigate, etc.

← Academic theme templates!

alison.rbind.io

Learn: alison.rbind.io/post/2017-06-12-up-and-running-with-blogdown

Bookdown books

Organize and navigate html files as e-books

The screenshot shows the 'R for Data Science' website. On the left is a table of contents with sections like 'Welcome', '1 Introduction', 'II Wrangle', etc. The main content area displays the title 'R for Data Science' by Garrett Golemund and Hadley Wickham, followed by a 'Welcome' section. The welcome text describes the book's purpose: teaching data science with R, covering data structure, cleaning, visualization, and reproducible research. Below the text is a book cover for 'R for Data Science' by O'Reilly, featuring a green parrot and the subtitle 'VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA'.

Really powerful for organizing reports and documents.

I wish I could have written my PhD thesis is Bookdown

- Eg: github.com/benmarwick/huskydown

r4ds.had.co.nz
Learn: bookdown.org/yihui/bookdown

Simple slides

Create slides in a single RMarkdown file



Free and Open Source Software for Data Science

JJ. Allaire
rstudio::conf 2020
1/29/2020

Imagine re-creating presentations with updated data.

Text-based slide creation can be a powerful flow to think and outline.

Share presentations – and with a human-readable url!

rstudio.com/slides/rstudio-pbc
Learn: rmarkdown.rstudio.com/lesson-11

Xaringan slides

Create slides in a single RMarkdown file

Presentation Ninja

✂
with xaringan

Yihui Xie

2016/12/12 (updated: 2019-02-07)

1 / 38

Highlight your code

Why? It makes it more readable ••

```
---  
title: "armdeck"  
subtitle: "rstudio::conf 2019"  
author: "Alison Hill"  
date: "r Sys.Date()"*  
output:  
  xaringan::moon_reader:  
    nature:  
      highlightStyle: github  
      highlightLines: true  
---
```

The highlight style options are:

- arta, ascetic, dark, default, far, github, googlecode, idea, ir-black, magula, monokai, rainbow, solarized-dark, solarized-light, sunburst, tomorrow, tomorrow-night-blue, tomorrow-night-bright, tomorrow-night, tomorrow-night-eighties, vs, zenburn.

41 / 91

Incorporate powerful styling options from within R (without requiring knowledge of JavaScript, CSS, etc)

slides.yihui.org/xaringan

arm.rbind.io/slides/xaringan

Learn: above, and bookdown.org/yihui/rmarkdown/xaringan

14

Exploring missing values in naniar

Allison Horst
2020-05-19

1. Introduction

2. Meet the data

3. Initial NA counts & proportions

4. Visualizing NAs

5. Explore NA intersections

6. Missing relationships

7. Keep exploring

MEDS talk examples

Start Over

1. Introduction



Missing values*, indicated by (or coerced to) `NA` in R, are common in environmental data due to equipment malfunction, survey non-response, human error, resource limitations, and any number of other unforeseen hiccups that can occur during data collection. Despite their ubiquity, `NA`s are rarely considered in exploratory data analysis, and are commonly “dealt with” (read: disappeared) by listwise deletion. Listwise deletion (in which any row with an `NA` is removed) *may* be the best method for handling missings, but also omits valuable existing observations, reduces statistical power, and depending on the mechanism of missingness can increase bias in parameter estimates. Exploring and thinking critically about missing data is an important and often overlooked part of exploratory data analysis that can help us to understand **what** data are missing and **why**, so that we choose an appropriate method for handling them.

But **how do we explore and visualize data that don't exist?**

In this tutorial, we will move beyond `is.na()` to learn other useful tools and approaches for exploring and visualizing missing values with helpful functions in the `naniar` package by Dr. Nick Tierney.

* Here, I use “missing values” to describe any missing data record (`NA`), which can be any type (e.g. character, date, etc.) and does not imply only numeric data.

Reimagine teaching and how to blend lectures and hands-on coding for learners of all levels

allisonhorst.shinyapps.io/missingexplorer

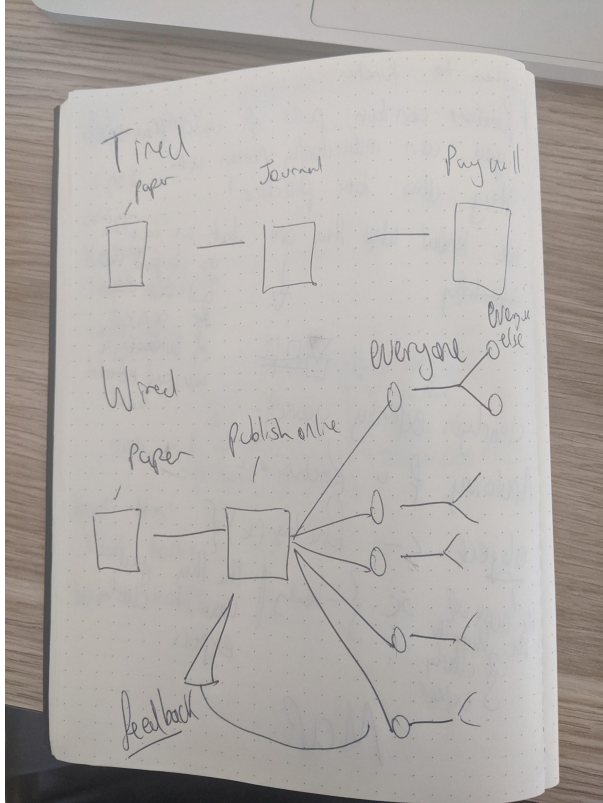
Learn: education.rstudio.com/blog/2020/05/learn-for-remote

Discussion time

What examples or questions do you have?

Possible discussion topics

- How does RMarkdown relate to/streamline the academic publishing process?
 - Analog: rOpenSci software review process
- Friendly entryways to open science & publishing : you're already doing it w/ code
 - Process affects the outcome: Easier to share at the end because you're already sharing with yourself throughout
- Not just R! Examples from other languages (Jupyter [note]books)
- Open publishing in the wild
 - Education: allisonhorst.github.io, datavizm20.classes.andrewheiss.com, tinystats.github.io/teacups-giraffes-and-statistics, ida.numbat.space
 - Programs: openscapes.org
 - Accompanying science pubs: ohi-science.org/betterscienceinlesstime



**You're an academic
who wants to spend the
least amount of time
formatting PDFs for
submitting to journals?
We've got you.**

Possible discussion topics

- RMarkdown <> Word workflows: noamross.github.io/redoc
- Nick's experience writing his thesis in bookdown: how does it compare to latex?
- Incorporating RMarkdown sub-documents (“knit child”): [OHI suppl. methods](#)
- How to share documents using GitHub's gh-pages or doc/: [R for Excel Users](#)