

R for the Excel User

Julie Lowndes & Allison Horst

2019-09-29

Contents

1	Welcome	7
1.1	Prerequisites	7
2	Overview	9
2.1	Summary (a few sentences)	9
2.2	Objectives (more detailed, bulletpoints?)	9
2.3	Resources	9
2.4	Overview	9
2.5	RStudio Orientation	12
2.6	R Scripts	15
2.7	Don't save the workspace	16
2.8	Deep thought: keep the raw data raw.	16
2.9	Interludes (deep thoughts/openscapes)	16
2.10	Our Turn Your Turn 1	16
2.11	Our Turn Your Turn 2	16
2.12	Efficiency Tips	16
3	readxl	17
3.1	Summary (a few sentences)	17
3.2	Objectives (more detailed, bulletpoints?)	17
3.3	Resources	18
3.4	Lessons teaching for each objective..... (objectives, examples) . .	18
3.5	Fun facts (quirky things) - making a note of these wherever possible for interest (little "Did you know?" sections)	19
3.6	Interludes (deep thoughts/openscapes)	19
3.7	Our Turn Your Turn 1	19
3.8	Our Turn Your Turn 2	19
3.9	Efficiency Tips	19
4	RMarkdown	21
4.1	Summary (a few sentences)	21
4.2	Objectives (more detailed, bulletpoints?)	21
4.3	Resources	21
4.4	Lessons teaching for each objective..... (objectives, examples) . .	21

4.5	Fun facts (quirky things) - making a note of these wherever possible for interest (little “Did you know?” sections)	21
4.6	Interludes (deep thoughts/openscapes)	21
4.7	Our Turn Your Turn 1	21
4.8	Our Turn Your Turn 2	21
4.9	Efficiency Tips	21
5	Dplyr and Pivot Tables	23
5.1	Summary (a few sentences)	23
5.2	Objectives (more detailed, bulletpoints?)	23
5.3	Resources	23
5.4	Lessons teaching for each objective..... (objectives, examples)	23
5.5	Fun facts (quirky things) - making a note of these wherever possible for interest (little “Did you know?” sections)	23
5.6	Interludes (deep thoughts/openscapes)	23
5.7	Our Turn Your Turn 1	23
5.8	Our Turn Your Turn 2	23
5.9	Efficiency Tips	23
6	Dplyr and vlookups	25
6.1	Summary (a few sentences)	25
6.2	Objectives (more detailed, bulletpoints?)	25
6.3	Resources	25
6.4	Lessons teaching for each objective..... (objectives, examples)	25
6.5	Fun facts (quirky things) - making a note of these wherever possible for interest (little “Did you know?” sections)	25
6.6	Interludes (deep thoughts/openscapes)	25
6.7	Our Turn Your Turn 1	25
6.8	Our Turn Your Turn 2	25
6.9	Efficiency Tips	25
7	Tidying	27
7.1	Better practices [needs a better name]	27
7.2	Summary (a few sentences)	27
7.3	Objectives (more detailed, bulletpoints?)	27
7.4	Resources	28
7.5	Lessons teaching for each objective..... (objectives, examples)	28
7.6	Fun facts (quirky things) - making a note of these wherever possible for interest (little “Did you know?” sections)	29
7.7	Interludes (deep thoughts/openscapes)	29
7.8	Our Turn Your Turn 2	29
7.9	Efficiency Tips	29
8	Formatting and Sharing	31
8.1	Summary (a few sentences)	31
8.2	Objectives (more detailed, bulletpoints?)	31

8.3	Resources	31
8.4	Lessons teaching for each objective..... (objectives, examples) . .	31
8.5	Fun facts (quirky things) - making a note of these wherever possible for interest (little “Did you know?” sections)	32
8.6	Interludes (deep thoughts/openscapes)	32
8.7	Our Turn Your Turn 1	32
8.8	Our Turn Your Turn 2	32
8.9	Efficiency Tips	32
9	Synthesis	33
9.1	Summary (a few sentences)	33
9.2	Objectives (more detailed, bulletpoints?)	33
9.3	Resources	33
9.4	Lessons teaching for each objective..... (objectives, examples) . .	33
9.5	Fun facts (quirky things) - making a note of these wherever possible for interest (little “Did you know?” sections)	33
9.6	Interludes (deep thoughts/openscapes)	33
9.7	Our Turn Your Turn 1	33
9.8	Our Turn Your Turn 2	33
9.9	Efficiency Tips	33

Chapter 1

Welcome

TODO: shorten, add about RStudio::conf

Excel is a widely used and powerful tool for working with data. As automation, reproducibility, collaboration, and frequent reporting become increasingly expected in data analysis, a good option for Excel users is to extend their workflows with R. Integrating R into data analysis with Excel can bridge the technical gap between collaborators using either software. R enables use of existing tools built for specific tasks and overcomes some limitations that arise when working with large datasets or repeated analyses. This course is for Excel users who want to add or integrate R and RStudio into their existing data analysis toolkit. Participants will get hands-on experience working with data across R, Excel, and Google Sheets, focusing on: data import and export, basic wrangling, visualization, and reporting with RMarkdown. Throughout, we will emphasize conventions and best practices for working reproducibly and collaboratively with data, including naming conventions, documentation, organization, all while “keeping the raw data raw”. Whether you are working in Excel and want to get started in R, already working in R and want tools for working more seamlessly with collaborators who use Excel, or whether you are new to data analysis entirely, this is the course for you!

1.0.1 About this book

1.1 Prerequisites

Before the training, please make sure you have done the following:

1. Download and install **up-to-date versions** of:
 - R: <https://cloud.r-project.org>
 - RStudio: <http://www.rstudio.com/download>

2. Install the Tidyverse
3. Get comfortable: if you're not in a physical workshop, be set up with two screens if possible. You will be following along in RStudio on your own computer while also watching a virtual training or following this tutorial on your own.

Chapter 2

Overview

2.1 Summary (a few sentences)

2.2 Objectives (more detailed, bulletpoints?)

2.3 Resources

R is not only a language, it is an active community of developers, users, and educators (often these traits are in each person). This workshop and book based on many excellent materials created by other members in the R community, who share their work freely to help others learn. Using community materials is how WE learned R, and each chapter of the book will have Resources listed for further reading into the topics we discuss. And, when there is no better way to explain something (ahem Jenny Bryan), we will quote or reference that work directly.

- What They Forgot to Teach You About R — Jenny Bryan & Jim Hester
- Stat545 — Jenny Bryan & Stat545 TAs
- Where do Things Live in R? REX Analytics

2.4 Overview

Welcome.

This workshop you will learn hands-on how to begin to interoperate between Excel and R.

A main theme throughout is to produce analyses people can understand and build from — including Future You. Not so brittle/sensitive to minor changes.

We will learn and reinforce X main things all at the same time: coding with best practices (R/RStudio), how this relates to operations in Excel, Z. This training will teach these all together to reinforce skills and best practices, and get you comfortable with a workflow that you can use in your own projects.

2.4.1 What to expect

This is going to be a fun workshop.

The plan is to expose you to X that you can have confidence using in your work. You'll be working hands-on and doing the same things on your own computer as we do live on up on the screen. We're going to go through a lot in these two days and it's less important that you remember it all. More importantly, you'll have experience with it and confidence that you can do it. The main thing to take away is that there *are* good ways to work between R and Excel; we will teach you to expect that so you can find what you need and use it! A theme throughout is that tools exist and are being developed by real, and extraordinarily nice, people to meet you where you are and help you do what you need to do. If you expect and appreciate that, you will be more efficient in doing your awesome science.

You are all welcome here, please be respectful of one another. You are encouraged to help each other.

Everyone in this workshop is coming from a different place with different experiences and expectations. But everyone will learn something new here, because there is so much innovation in the data science world. Instructors and helpers learn something new every time, from each other and from your questions. If you are already familiar with some of this material, focus on how we teach, and how you might teach it to others. Use these workshop materials not only as a reference in the future but also for talking points so you can communicate the importance of these tools to your communities. A big part of this training is not only for you to learn these skills, but for you to also teach others and increase the value and practice of open data science in science as a whole.

2.4.2 What you'll learn

TODO: dev

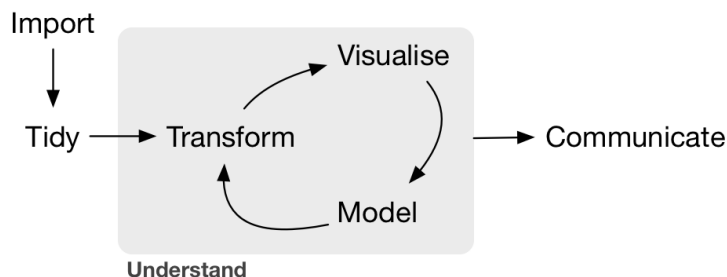
- Motivation is to bridge and/or get out of excel
- We're not going to replicate all of your fancy things in R,
- We use Excel to look at data that we're reading into R

- Spreadsheets are great; blend data entry with analyses and we're going to try to help you think about them a bit more distinctively.
- Most important collaborator is future you, and future us

An important theme for this workshop is being deliberate about your analyses and setting things up in a way that will make your analytical life better downstream in the current task, and better when Future You or Future Us revisit it in the future (i.e. avoiding: what happens next? What does this name mean?)

This graphic by Hadley Wickham and Garrett Golemund in their book R for Data Science is simple but incredibly powerful:

```
knitr::include_graphics("img/r4ds_data-science.png")
```



You may not have ever thought about analysis in such discrete steps: I certainly hadn't before seeing this. That is partly because in Excel, it can be easy to blend these steps together. We are going to keep them separate, and talk about why. The first step is Import: and implicit in this as a first step is that the data is stored elsewhere and is not manipulated directly, which **keeps the raw data raw**.

We will be focusing on:

- **Import:** `readr`, `readxl` to read raw data stored in CSV or Excel files directly into R
- **Tidy:** `tidyr` to (re)organize rows of data into unique values
- **Transform:** `dplyr` to “wrangle” data based on subsetting by rows or columns, sorting and joining
- **Visualize:** `ggplot2` static plots, using grammar of graphics principles
- **Communicate**
 - `writexl` to export intermediate and final data
 - GitHub File Upload and Issues for online publishing and collaboration

2.4.3 Emphasizing collaboration

TODO: rewrite/update (from OHI book):

Collaborating efficiently has historically been really hard to do. It's only been the last 20 years or so that we've moved beyond mailing things with the postal service. Being able to email and get feedback on files through track changes was a huge step forward, but it comes with a lot of bookkeeping and reproducibility issues (did I send that report based on `analysis_final_final.xls` or `analysis_final_usethisone.xls`?). But now, open tools make it much easier to collaborate.

Working with collaborators in mind is critical for reproducibility. And, your most important collaborator is Future You. This training will introduce best practices using open tools, so that collaboration will become second nature to you!

2.4.4 By the end of the course...

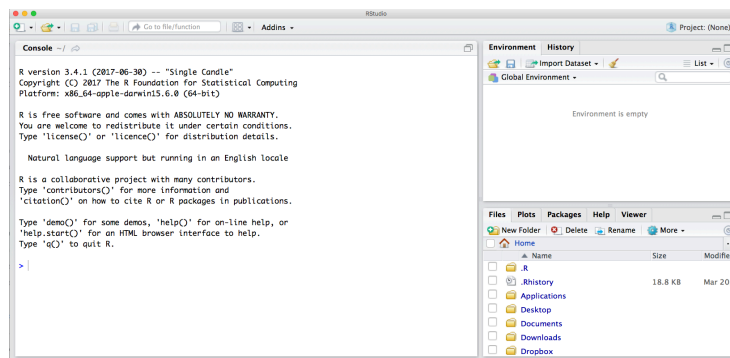
By the end of this course you'll produce this report that you can reproduce, which means... Introduce the problem we will solve. Eg: (just an idea maybe time-series is not a great idea) SMALL PROBLEM. (4 mins) Show data files, We will discuss our analysis plan (only enough to motivate!) Create a report, that looks great.

2.5 RStudio Orientation

Open RStudio for the first time.

Launch RStudio/R.

```
knitr::include_graphics("img/RStudio_IDE.png")
```



Notice the default panes:

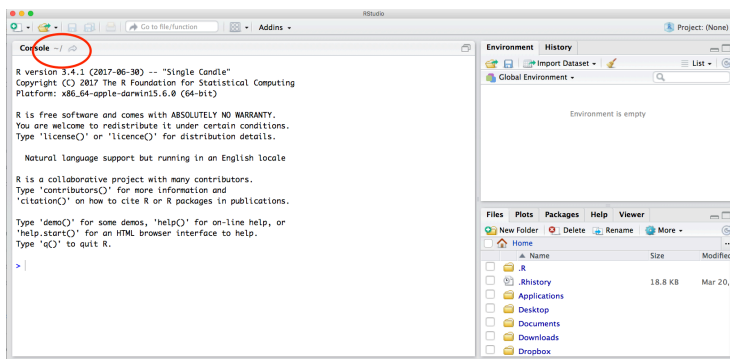
- Console (entire left)
- Environment/History (tabbed in upper right)
- Files/Plots/Packages/Help (tabbed in lower right)

FYI: you can change the default location of the panes, among many other things: Customizing RStudio.

An important first question: **where are we?**

If you've have opened RStudio for the first time, you'll be in your Home directory. This is noted by the `~/` at the top of the console. You can see too that the Files pane in the lower right shows what is in the Home directory where you are. You can navigate around within that Files pane and explore, but note that you won't change where you are: even as you click through you'll still be Home: `~/`.

```
knitr::include_graphics("img/RStudio_IDE_homedir.png")
```



2.5.1 RStudio Projects

Create a new Project called 'r-for-excel-users'. File > New Project... > New Directory > New Project. Give your Project a name browse to a place to keep it. And then click to Create Project!

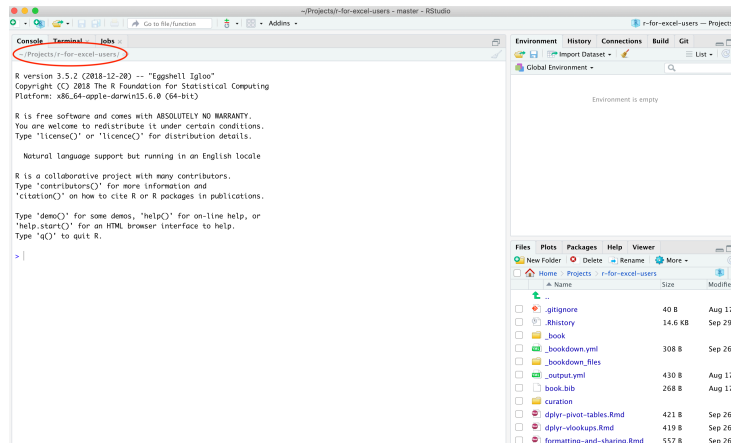
What is a Project? It is a way to organize all of the relevant things you need for an analysis in the same place. This means data, code, figures, notes, etc.

Why does this matter? Keeping everything you need for your analysis together makes it less brittle and more portable — across people, time, and computers.

Working directory = no file path/broken path issues. Notice that a folder now appears wherever you saved this project with the same name, and it contains a `.Rproj` file.

Now that we have our Project, here is an important question: where are we? Now we are in our Project. Everything we do will by default be saved here so we can be nice and organized.

```
knitr::include_graphics("img/RStudio_IDE_projdir.png")
```



2.5.2 R Console

Watch me work in the Console.

I can do math:

```
52*40
365/12
```

But like Excel, the power comes not from doing small operations by hand (like $8*22.3$), it's by being able to operate on swaths of numbers. Columns, Rows, etc. In Excel, they are stored in the spreadsheet. In R, they are stored in variables.

R stores data in variables, and that data can be a variety of formats, like numeric and text. Let's have a look at some data in R. R has several built-in data sets that we can look at and work with.

One of these datasets is called `mtcars`. If I write this in the Console, it will print the data in the console.

```
mtcars
```

I can also use RStudio's Viewer to see this in a more familiar-looking format:

```
View(mtcars)
```

This opens the fourth pane of the RStudio IDE. In the Viewer I can do things like filter or sort. This does not do anything to the actual data, it just changes how you are viewing the data.

There are also functions like in Excel; I can do `?sum` and look at the help pages.

Average mile per gallon. In excel there is a function called AVERAGE. Let's see

OK so working in the Console is great, but it gets messy. It's good for quick things. But looking back at what I've done and trying to build upon it would be a nightmare.

2.5.3 Deep thought: Error messages are your friends

Implicit contract with the computer / scripting language: Computer will do tedious computation for you. In return, you will be completely precise in your instructions. Typos matter. Case matters. Pay attention to how you type.

Remember that this is a language, not unsimilar to English! There are times you aren't understood – it's going to happen. There are different ways this can happen. Sometimes you'll get an error. This is like someone saying 'What?' or 'Pardon'? Error messages can also be more useful, like when they say 'I didn't understand what you said, I was expecting you to say blah'. That is a great type of error message. Error messages are your friend. Google them (copy-and-paste!) to figure out what they mean.

—>

And also know that there are errors that can creep in more subtly, when you are giving information that is understood, but not in the way you meant. Like if I am telling a story about suspenders that my British friend hears but silently interprets in a very different way (true story). This can leave me thinking I've gotten something across that the listener (or R) might silently interpreted very differently. And as I continue telling my story you get more and more confused... Clear communication is critical when you code: write clean, well documented code and check your work as you go to minimize these circumstances!

2.6 R Scripts

Instead of working in the Console that can get messy, we can be more organized. In a script. Let's all do this together.

File > New File > R Script.

This is a blank slate for us to write our code; but there are some good practices we can start off with. Let's add a useful header to the top of it. For example, at a minimum:

```
# -----
# A descriptive title
# Summary of what this script is for
# Your name
# Contact information
# Date
```

```
# -----  
  
# Other things you might include: required packages or datasets, relevant links (e.g.
```

Since we're working in or Project, this script is now nicely saved in our Project.

Let's attach a package. Since you've already installed tidyverse

```
# Attach the tidyverse  
library(tidyverse)
```

Cool looking plot with diamonds.

You can run code line by line or sourcing the script

2.7 Don't save the workspace

2.8 Deep thought: keep the raw data raw.

Discussing using Excel for variables.

Horror Stories! Economist etc.

2.9 Interludes (deep thoughts/openscapes)

Comments! Organization (spacing, subsections, vertical structure, indentation, etc.)! Well-named variables! Also, well-named operations so analyses (select(data, columnname)) instead of data[1:6,5] and excel equivalent. (Ex with strings) Not so brittle/sensitive to minor changes.

2.10 Our Turn Your Turn 1

2.11 Our Turn Your Turn 2

2.12 Efficiency Tips

Chapter 3

readxl

Note for Allison: the `{#readxl}` above lets us reference it from another chapter...not sure how much we'll use it but it's possible (and doesn't hurt to have). Cool, thanks Julie!

3.1 Summary (a few sentences)

The **readxl** package makes it easy to import tabular data from Excel spreadsheets (.xls or .xlsx files) and includes several options for cleaning data during import. **readxl** has no external dependencies and functions on any operating system, making it an OS- and user-friendly package that simplifies getting your data from Excel into R.

3.2 Objectives (more detailed, bulletpoints?)

- Use `readr::read_csv()` to read in a comma separated value (CSV) file
- Use `readxl::read_excel()` to read in an Excel worksheet
- Use `readxl::read_excel()` to read in parts of a worksheet (by cell range)
- Specify column names when importing Excel data
- Replace a specific string/value in a spreadsheet with with 'NA'
- Skip *n* rows when importing an Excel worksheet
- Read in multiple Excel worksheets
- Something about readxl workflows (e.g. `%>% write_csv()` to keep a copy of the raw data in a project)??

3.3 Resources

- <https://readxl.tidyverse.org/>

3.4 Lessons teaching for each objective..... (objectives, examples)

3.4.1 Prep: get data files into your working directory

In Session 1, we introduced how and why R Projects are great for reproducibility, because our self-contained working directory will be the **first** place R looks for files.

You downloaded four files for this workshop:

- fish_counts_curated.csv
- invert_counts_curated.xlsx
- kelp_counts_curated.xlsx
- ...the OTHER ONE?

Copy and paste those files into the ‘RforExcelUsers’ folder on your computer. Notice that now these files show up right in your working directory when you go back to that Project in RStudio! That means they’re going to be in the first place R will look when you ask it to find a file.

3.4.2 Create a new .R script and attach the tidyverse

In your RforExcelUsers project in RStudio, open a new .R script and add a useful header to the top of it. For example, at a minimum:

```
# -----
# A descriptive title
# Summary of what this script is for
# Your name
# Contact information
# -----

# Other things you might include: required packages or datasets, relevant links (e.g.
```

We’ll read in a CSV file with `readr::read_csv()`, so we need to have the `readr` package attached. Since it’s part of the `tidyverse`, we’ll go ahead and attach the `tidyverse` package below our script header using `library(package_name)`.

```
# Attach the tidyverse
library(tidyverse)
```

3.5. FUN FACTS (QUIRKY THINGS) - MAKING A NOTE OF THESE WHEREVER POSSIBLE FOR INTEREST

Now, all of the packages and functions within the `tidyverse`, including `readr::read_csv()` are available for us to use.

3.4.3 Use `readr::read_csv()` to read in data from a CSV file

There are many types of files containing data that you might want to work with in R. A common one is a comma separated value (CSV) file, which contains values with each column entry separated by a comma delimiter. CSVs can be opened, viewed, and worked with in Excel just like an .xls or .xlsx file - but let's learn how to get data directly from a CSV into R where we can work with it more reproducibly.

The CSV we'll read in here is called "fish_counts_curated.csv", and contains observations for "the abundance and size of fish species as part of SBCLTER's kelp forest monitoring program to track long-term patterns in species abundance and diversity" from the Santa Barbara Channel Long Term Ecological Research program.

Source: Reed D. 2018. SBC LTER: Reef: Kelp Forest Community Dynamics: Fish abundance. Environmental Data Initiative. <https://doi.org/10.6073/pasta/dbd1d5f0b225d903371ce89b09ee7379>. Dataset accessed 9/26/2019.

Read in the "fish_counts_curated.csv"

3.5 Fun facts (quirky things) - making a note of these wherever possible for interest (little "Did you know?" sections)

3.6 Interludes (deep thoughts/openscapes)

3.7 Our Turn Your Turn 1

3.8 Our Turn Your Turn 2

3.9 Efficiency Tips

Chapter 4

RMarkdown

- 4.1 Summary (a few sentences)
- 4.2 Objectives (more detailed, bulletpoints?)
- 4.3 Resources
- 4.4 Lessons teaching for each objective..... (objectives, examples)
- 4.5 Fun facts (quirky things) - making a note of these wherever possible for interest (little “Did you know?” sections)
- 4.6 Interludes (deep thoughts/openscapes)
- 4.7 Our Turn Your Turn 1
- 4.8 Our Turn Your Turn 2
- 4.9 Efficiency Tips

Chapter 5

Dplyr and Pivot Tables

- 5.1 Summary (a few sentences)
- 5.2 Objectives (more detailed, bulletpoints?)
- 5.3 Resources
- 5.4 Lessons teaching for each objective..... (objectives, examples)
- 5.5 Fun facts (quirky things) - making a note of these wherever possible for interest (little “Did you know?” sections)
- 5.6 Interludes (deep thoughts/openscapes)
- 5.7 Our Turn Your Turn 1
- 5.8 Our Turn Your Turn 2
- 5.9 Efficiency Tips

Chapter 6

Dplyr and vlookups

- 6.1 Summary (a few sentences)
- 6.2 Objectives (more detailed, bulletpoints?)
- 6.3 Resources
- 6.4 Lessons teaching for each objective..... (objectives, examples)
- 6.5 Fun facts (quirky things) - making a note of these wherever possible for interest (little “Did you know?” sections)
- 6.6 Interludes (deep thoughts/openscapes)
- 6.7 Our Turn Your Turn 1
- 6.8 Our Turn Your Turn 2
- 6.9 Efficiency Tips

Chapter 7

Tidying

7.1 Better practices [needs a better name]

How to be a nimble useR Modern useRs are nimble internet useRs something clever about cleaning I am the worst at naming things

7.2 Summary (a few sentences)

R ecosystem evolves and improves due to contributed work by the community, and this is a good thing. Being a nimble useR means being able to navigate/keep tabs on this ecosystem and find what you need. It also means working reproducibly, so you can re-run and update things more easily. Here we will teach you how to expect things and help yourself. Pay attention to urls.

7.3 Objectives (more detailed, bulletpoints?)

- expect there is a better way, how and where to look (20 mins)
 - CRAN
 - Twitter #rstats
 - rOpenSci
 - RStudio
 - Example: how to Google.
- hands-on with janitor (30+ mins)
 - discovery and quality assurance
 - installing from GitHub
 - big payoff for little effort

- hands-on with another excel-useful example: `skimr`?
- reproducibility (20 mins)
 - it's important, scripted

7.4 Resources

- Wilson et al. 2014 “Good enough practices”

7.5 Lessons teaching for each objective..... (objectives, examples)

7.5.1 Expect there's a better way chat

- give time for them to google?

7.5.2 Janitor

janitor & other things that will make your life easier with limited effort Janitor: up till now the column names have been fine. Until now.

7.5.2.1 Our turn your turn

Walk through and example and leave our code up, and have you do it but clean another dataset. Work with a neighbor.

7.5.3 Example: How to Google

Pay attention to URLs, build github/rmarkdown savviness (ex: `raw.githubusercontent.com`)

- I read this blog: <https://blog.revolutionanalytics.com/2018/08/how-to-use-r-with-excel.html>
- I've never heard of click on `openxlsx`, what is it
- Takes me here <https://www.rdocumentation.org/packages/openxlsx/versions/4.1.0.1>, but I want more info. How recently was it worked on? Does it interface with tidyverse? Click on “news”
- Takes me here. <https://raw.githubusercontent.com/awalker89/openxlsx/master/NEWS> . Not useful. But from this URL, - I see the username so I can edit this url to be <https://github.com/awalker89/openxlsx/>

7.6. *FUN FACTS (QUIRKY THINGS) - MAKING A NOTE OF THESE WHEREVER POSSIBLE FOR INTEREST*

- 1st thing: most recent commit was a year ago. Can poke around more, are there issues open, are they taken care of? Etc. I will probably not pursue using this right now. But good to have learned about it.

7.6 Fun facts (quirky things) - making a note of these wherever possible for interest (little “Did you know?” sections)

7.7 Interludes (deep thoughts/openscapes)

7.8 Our Turn Your Turn 2

7.9 Efficiency Tips

- browser efficiency tips
 - Rmd/github anchors for urls
 - press command to open a new tab

Reproducibility is important (this might be new to some people) Example: run everything start to finish and then closing it all and trying to do again In excel Vs R If your computer shuts off are you nervous to close it? Recreate it “What they didn’t forget to teach you about R” WTDF. uncool

Chapter 8

Formatting and Sharing

8.1 Summary (a few sentences)

8.2 Objectives (more detailed, bulletpoints?)

8.3 Resources

8.4 Lessons teaching for each objective..... (objectives, examples)

1. Create a GitHub account: <https://github.com> *Note! Shorter names that kind of identify you are better, and use your work email!*

- 8.5 Fun facts (quirky things) - making a note of these wherever possible for interest (little “Did you know?” sections)
- 8.6 Interludes (deep thoughts/openscapes)
- 8.7 Our Turn Your Turn 1
- 8.8 Our Turn Your Turn 2
- 8.9 Efficiency Tips

Chapter 9

Synthesis

- 9.1 Summary (a few sentences)
- 9.2 Objectives (more detailed, bulletpoints?)
- 9.3 Resources
- 9.4 Lessons teaching for each objective..... (objectives, examples)
- 9.5 Fun facts (quirky things) - making a note of these wherever possible for interest (little “Did you know?” sections)
- 9.6 Interludes (deep thoughts/openscapes)
- 9.7 Our Turn Your Turn 1
- 9.8 Our Turn Your Turn 2
- 9.9 Efficiency Tips