Software Carpentry MBARI - Nov-Dec 2017

Please feel free to write questions here, or answer others' questions too. You can also use this doc for shared notes.

https://github.com/OHI-Science/data-science-training

Content:

Day 2 - make sure you have stickies!

The training book website is down :(.

We can look at the book from another url, here. (if you're curious, this uses RawGit.com, which will display rendered html).

Collaborate on Github

- https://jules32.github.io/collaborate/collab.html

- The user1's collaborate repo.
- User1 New repository: https://github.com/user1/collaborate/
    - user1: Settings tab -> Collaborators: add user2
    - user2: accept invitation at https://github.com/user1/collaborate/invitations

Final Topics Suggested

- Knit to presentations, websites, etc
    - Rmd Formats
        - Website
            - R Markdown Websites - Rmd Lesson 13
            - R Markdown **Websites**
            -
        - Presentation
            - Slide Presentations - Rmd Lesson 11
            - Presentations with **ioslides**
        - Advanced R Markdown
    - Latest advanced R packages
        - Presentation: **xaringan**: Presentation Ninja
        - Website: **blogdown**: Create Blogs and Websites with R Markdown
    - http://rmarkdown.rstudio.com/rmarkdown_websites.html
- Mapping, interactive visualization, shiny
    - Interactive Documents - Rmd Lesson 14
        - htmlwidgets for R
    - Visualization in R: plots & maps, static & interactive
- Tibbles
    - tibbles - R for Data Science
    - many models - R for Data Science
    - TrelliscopeJS: amazing lattice visualization of static or interactive plots, many at once
- How to use dplyr without knowing SQL on a database?
    - **dbplyr**: introduction
    - SQL databases and R - Data Carpentry R Ecology Lesson 5
- Using tidy style (dplyr/tidyr) on spatial (sf) and scientific (netcdf) data formats
    - **sf**: Simple Features for R: vector (points, lines, polygons)
        - Tidy spatial data in R: using dplyr, tidyr, and ggplot2 with sf
        - Spatial Data in R: New Directions
    - netcdf: **tidync**: A Tidy Approach to 'NetCDF' Data Exploration and Extraction

- Incorporating references into R Markdown
    - bbest/rmarkdown-example: R Markdown example showing figures & tables with captions, equations, inline R values and references with a Zotero library
    - Authorea.com: Google docs for academic writing
    - Rmd Bibliographies and Citations
    - crsh/citr: RStudio Addin to Insert Markdown Citations
- Community learning
    - eco-data-science.github.io
    - ecoquants.com/service/training/#community
    - Webinars – RStudio
- May the force be with you!
    - starlogs.net

Fun code for packages - will install if package isn't installed

```
pkgTest <- function(x)
{
  if (!require(x,character.only = TRUE))
  {
    install.packages(x,dep=TRUE)
    if(!require(x,character.only = TRUE)) stop("Package not found")
  }
}
```

Programming

Programming example:
https://github.com/OHI-Science/data-science-training/blob/master/programming.R

- Starting with lines 1-14: https://github.com/OHI-Science/data-science-training/blob/master/programming.R#L1-L14
    - Paste into File > New File > R Script
- Debugging in RStudio
    - https://support.rstudio.com/hc/en-us/articles/205612627-Debugging-with-RStudio
- Example up to function and for loop

```
# libraries ----
library(tidyverse)

# data ----
url <-
'https://raw.githubusercontent.com/OHI-Science/data-science-training/master/data/gapminder.csv'
```

```
gapminder <- read_csv(url) # View(gapminder)

# function: ----
country_plot <- function(cntry){ # cntry <- "Canada"

  #browser()
  png <- paste0("gdp_", cntry, ".png")
  cat("country_plot(", cntry, ") -> ", png, "\n")

  g <- gapminder %>%
    filter(country == cntry) %>%
    ggplot(aes(x = year, y = gdpPercap)) +
    geom_point() +
    geom_smooth() +
    labs(title = cntry)

  ggsave(png, g)
}
#country_plot("Mexico")

countries <- c("United States", "Peru", "New Zealand")
for (k in countries){
#for (k in 1:length(countries)){ # k=1
  cat(k, "\n")
  #cat(k,":", countries[k], "\n")
  #cat(sprintf("%03d: %s\n", k, countries[k]))

  country_plot(k)
  #country_plot(countries[k])
}


# new loop over ALL countries
countries <- unique(gapminder$country)

for (k in countries){
  cat(k, "\n")
  country_plot(k)
}
```

- If else

```
dir.create("developed")
dir.create("developing")

is_developed <- function(cntry, threshold=12000){ # cntry <- "Peru"
  gapminder %>%
    filter(country == cntry) %>%
    summarise(
      mean_gdp = mean(gdpPercap)) %>%
    .$mean_gdp >= threshold
}
is_developed("Peru", threshold = 2000)
is_developed("United States")

for (k in countries){
  cat(k, "\n")
  country_plot(k)
}
```

- Full deal: https://github.com/bbest/ds-repo/blob/master/plot_countries.R
- Quick Intro to Package Development with devtools - bbest Env Info course
-

OR, you can see the materials on GitHub itself (woop!). You can navigate to the Rmd files by going to:

github.com > ohi-science > data-science-training >

- Wrap up dplyr, tidyr. > tidyr.Rmd
- Programming. > programming.Rmd
- Collaborate with GitHub. > collaborating.Rmd
- TBD

## read in CO2 emissions data
co2 <- re
ad_csv("https://raw.githubusercontent.com/OHI-Science/data-science-training/master/data/co2.csv")

Gapminder wide dataset:
https://github.com/OHI-Science/data-science-training/blob/master/data/gapminder_wide.csv

Read into R with this code:

```
## wide format
gap_wide <-
readr::read_csv('https://raw.githubusercontent.com/OHI-Science/data-science-training/master/data/gapminder_wide.csv')
```

```
## yesterdays data format
gapminder <-
readr::read_csv('https://raw.githubusercontent.com/OHI-Science/data-science-training/master/data/gapminder.csv')
```

Code for wide to long format with gather and separate

~~~
## separate out the year
```{r}
gap_long <- gap_wide %>%
  gather(key = obstype_year,
         value = obs_values,
         -continent, - country) %>%
  separate(obstype_year,
           into = c("obs_type", "year"),
           sep = "_",
           convert = T)

head(gap_long)
```
~~~

Plotting life exp for americas

~~~
```{r}
life_df <- gap_long %>%
  filter(obs_type == "lifeExp",
         continent == "Americas")

head(life_df)
~~~

~~~~
ggplot(data = life_df, aes(x = year, y = obs_values, color = country)) +
  geom_line()
```
~~~~

7.7 Exercise

1. Using `gap_long`, calculate and plot the the mean life expectancy for each continent over time from 1982 to 2007. Give your plot a title and assign x and y labels. Hint: use the `dplyr::group_by()` and `dplyr::summarize()` functions.

7.8 Exercise

1. Convert “gap_long” all the way back to gap_wide. Hint: you’ll need to create appropriate labels for all our new variables (time*metric combinations) with the opposite of separate: `tidyr::unite()`.

Current Status

- Rmd code: aes(x = displ, y = hwy
    - Rendered web page: https://bbest.github.io/ds-repo/
- Julie’s repo: https://github.com/jules32/ds-repo
- Julie’s website: https://jules32.github.io/ds-repo/
-

Questions

- Amanda: What’s up with the assignment operator `<-` vs simply `=`?
    - Ben: The assignment operator in R can be either `<-` or `=`, but <- is preferred according to The tidyverse style guide, which we also recommend. Historically, S allowed assignment to the left `a <- 5` or to the right `5 -> a`. This is still possible in R, but again not recommended. So the `<-` is a carryover of this functionality. And the `=` tends to be associated with assignment of argument values, eg `seq(from = 1, to = 10)`. [Amanda: Thanks Ben!]

    -
- Heather: What's the difference between R Markdown and R Notebook? When would you use one vs the other?f
    - Ben: I'm not honestly totally clear about this, since they seem pretty much the same to me. You get either option under File > New File. You can read more about R Notebook here: rmarkdown.rstudio.com/r_notebooks.html which says, "An R Notebook is an R Markdown document with chunks that can be executed independently and interactively, with output visible immediately beneath the input." And check out all the Rmarkdown output formats at rmarkdown.rstudio.com/formats.html. [Thanks!!]
- In RMarkdown, what happens to the information you include after the letter r for example:

~~~
```{r}
```
~~~

Verse

~~~
```{ r is this r code or comments and does it show up?}
```
~~~

Thanks!

- Robin: ```{r this_is_the_name_of_your_code_chunk}```
    - OFten, each code chunk does a specific task, and the name of your code chunk can reflect that task (e.g., simulate_data)
    - http://kbroman.org/knitr_knutshell/pages/Rmarkdown.html
    - These code chunks get saved in a separate directory (called a 'cache'), and each code chunk is saved as a separate file (this speeds up the 'knitting' of your Rmd file to html, pdf, whatever)

- Can you do citations in Rmarkdown (like endnote?)
    - Yes! Here is an example of how: http://rmarkdown.rstudio.com/authoring_bibliographies_and_citations.html
    - Julie: we have used Zotero to collaboratively manage our references, export as a .bib file, and call in RMarkdown. Example in action: ohi-science/betterscienceinlesstime
- How do I resolve errors with git config --global?
    - If you're on a Mac, get git: https://git-scm.com/

Git config errors:

error key does not contain a section --global terminal

fatal: not in a git directory

- What about committing existing code (e.g. our test.rmd document) to a repo?
  - We'll do this next! http://ohi-science.org/data-science-training/github.html#sync-from-rstudio-to-github

- Is there a functional difference betweening using require() and library() for calling packages?
  - Elliott: require buries the error as a warning (for use in a function) while library will crash.

- **Dale**: Are there alternate style sheets that can be applied to the markdown to change the way it looks?
  - Ben: Rmarkdown - HTML Documents - Appearance and Style: note Bootswatch themes and CSS styling. For advanced templating, check out R Markdown Custom Formats.
- Julia: Git is asking me for my username and password every time I push. Did I miss a configuration/authentication step?
  - Ben: the storing of a password varies depending on platform (Windows, Mac, Linux). I recommend trying to do a push from the command line using the Git Bash shell (Windows) or Terminal (Mac). More gory details here: Caching your GitHub password in Git - User Documentation
-

Suggestions

- …

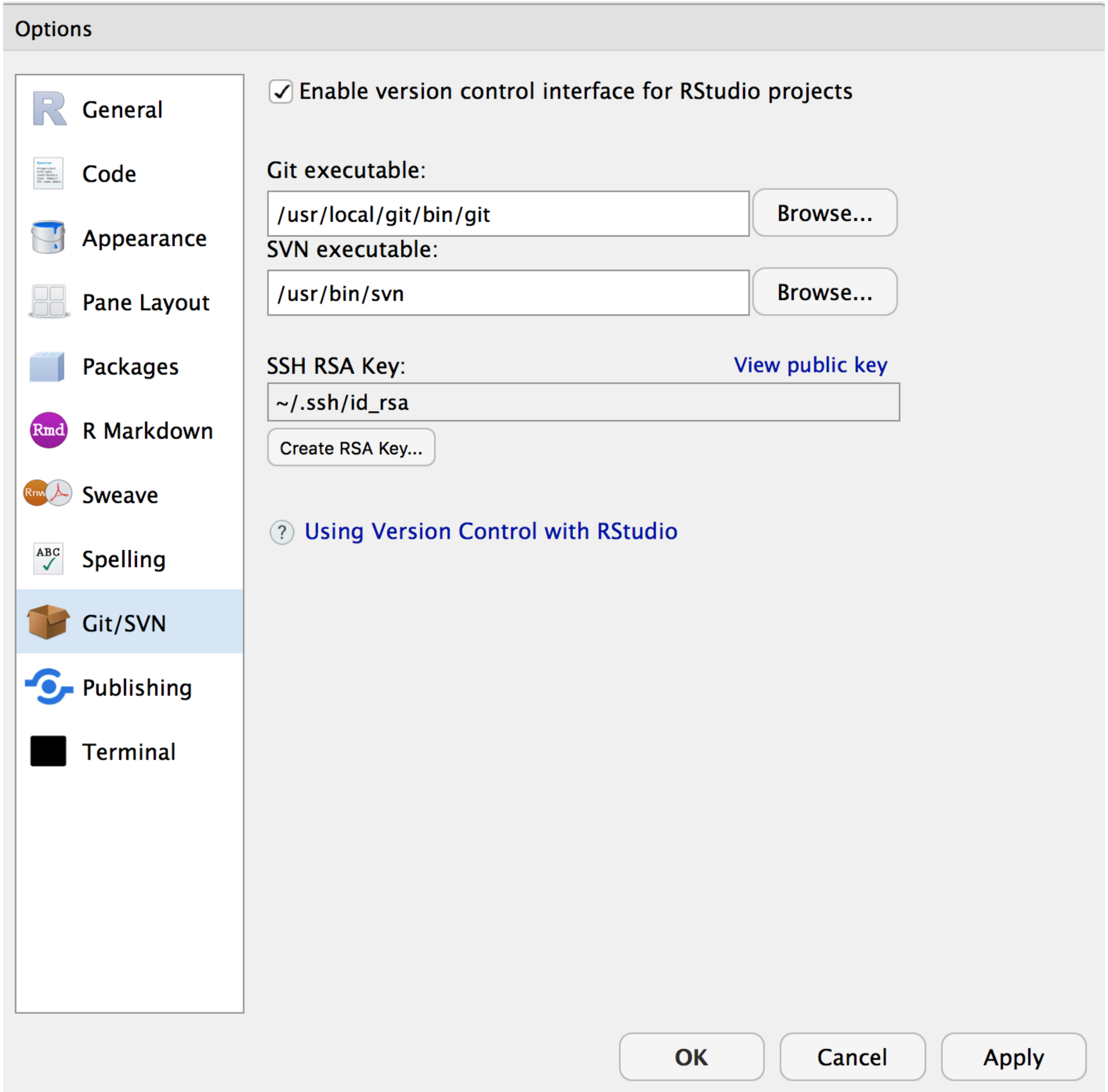
Git config errors (on a Mac)

error key does not contain a section --global terminal

fatal: not in a git directory

To solve:
In terminal: $ which git
Look at that url
In RStudio: Tools > Global Options > Git/SVN
Does the "Git executable" filepath match what the url in Terminal says? If not, click the browse button and navigate there.
Note: even though I navigated to /usr/local/bin/git, it is showing up as it does in the image below because /usr/local/bin/git was an alias on my computer. That is fine. Click OK.
Close RStudio

Reopen RStudio
Try cloning, and that should work and then you don't need to worry about typing into the Terminal, you're all done!



Options

☑ Enable version control interface for RStudio projects

Git executable:
/usr/local/git/bin/git

SVN executable:
/usr/bin/svn

SSH RSA Key:
~/.ssh/id_rsa

Resources

- Translating from Matlab to R:
    - R for MATLAB users: table of commands compared
    - matconv: R package to convert matlab code
    - R.matlab: R package to read and write MAT files and call MATLAB from within R

- What is R Markdown? 1 min video
- Rmarkdown.rstudio.com

    -
      ```
      output:
        html_document:
          code_folding: hide
          toc: true
          toc_float: true
      ```
    -
    -

To improve the lesson

GitHub Purpose:

1. Publish as webpages
1. Version your work
2. Offsite archive of your work

ohi-science.org/data-science-training is a website. github.com/ohi-science/data-science-training is the github repository.

Add Git errors from suggestions above
First commit: .gitignore and .Rproj

Move tidyverse figure to ggplot lesson