Research Tools

Michael Noonan

Biol 520C: Statistical modelling for biological data

Table of contents

Housekeeping

Housekeeping

- Great job on the talks.
- Term papers are due on Sunday (Dec. 12th), I will be submitting your grades on (Dec. 19th). Any assignments still missing by then will receive a grade of 0.
- Course/Instructor evaluations are due on Dec. 10th. Please take some time to fill them out.

R Markdown

R Markdown

R Markdown allows you to do two main things:

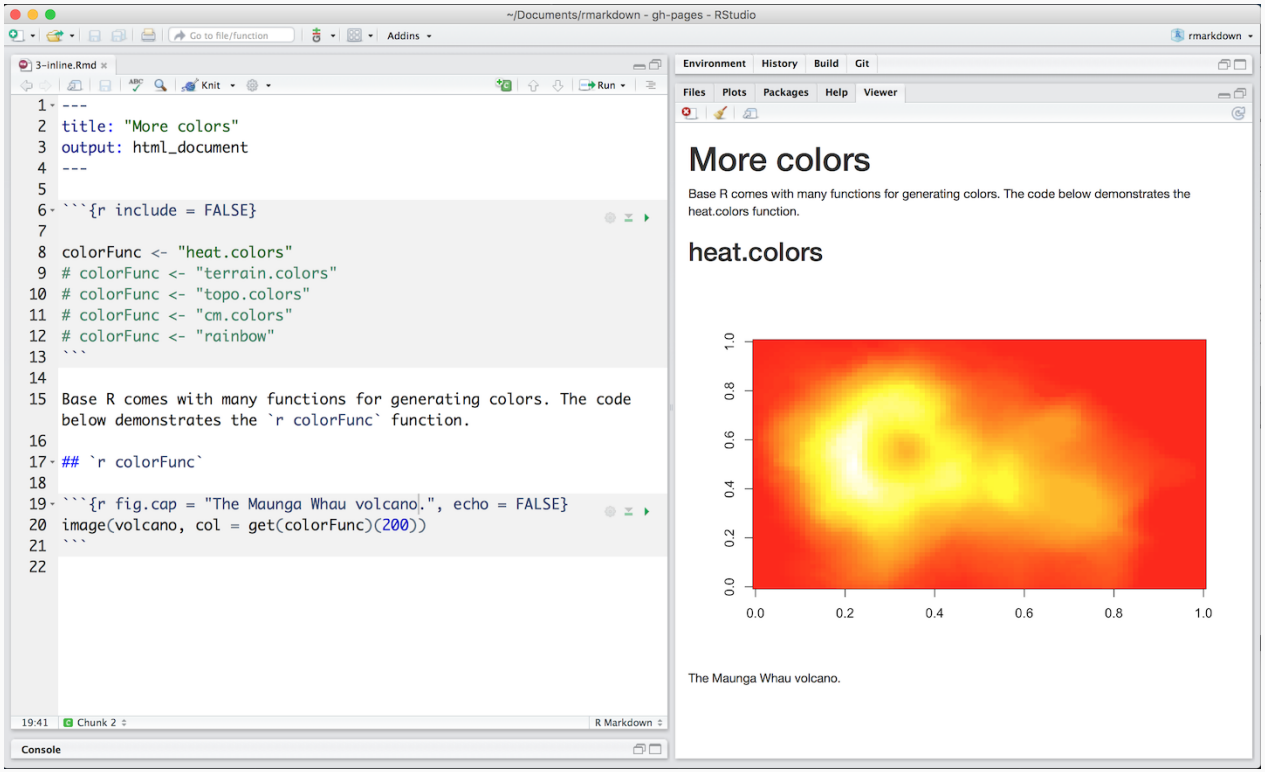
- save and execute code (how we have been using it in this course); and
- generate high quality reports that can be shared with an audience.

Importantly, you can do both of these things in a single R Markdown file.

This allows for easier reproducibility, since both the computing code and narratives are in the same document, and results are automatically generated from the source code/data.

Inline Code

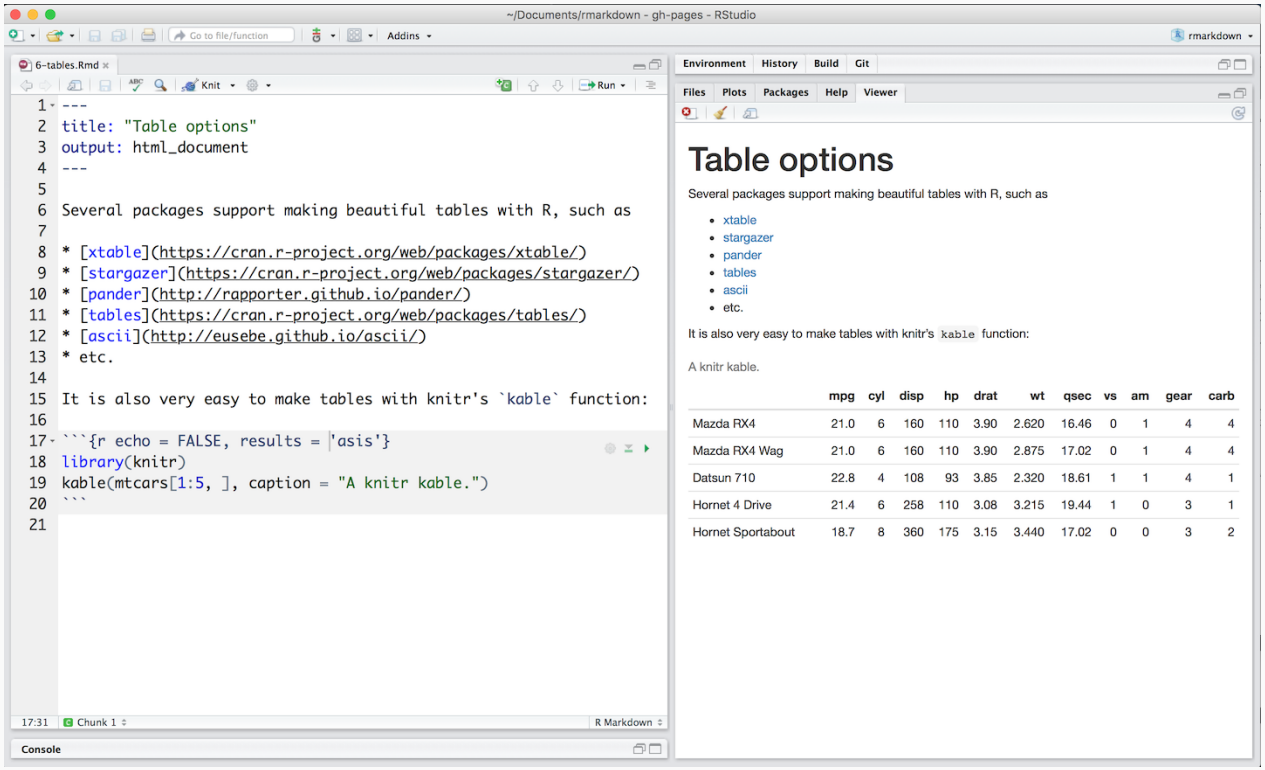
A useful feature of R Markdown is that code results can be inserted directly into the text of a .Rmd file.



This makes updating your documents easier, and prevents transcription errors.

Figures and tables

Figures and tables can also be generated directly from your data and places into your document.



Commenting

Just like using in R scripts to add comments, you can wrap text in the following: <!-- --> to add comments to your narratives that won't show up in the knit document.

R Markdown Summary

R Markdown is free, open source, and allows you to house all of your work (data cleaning and analysis, figures, tables, writing) in a single document.

It has built in spellchecking, commenting, and reference formatting and renders high quality reports on par with many other word processing software (e.g. Microsoft word, Pages, LaTeX).

Allows you to generate publication quality documents that are robust to issues like transcription errors and file disorganisation, ensuring reproducibility.

Also allows you to build web pages (e.g., the 520C course website was built entirely using R Markdown).

R Markdown Help

If you ever have any questions about how to do something in R Markdown, the R Markdown Definitive Guide is a good place to look:

https://bookdown.org/yihui/rmarkdown/

You can also access the cheatsheet from the help tab in `R Studio`.

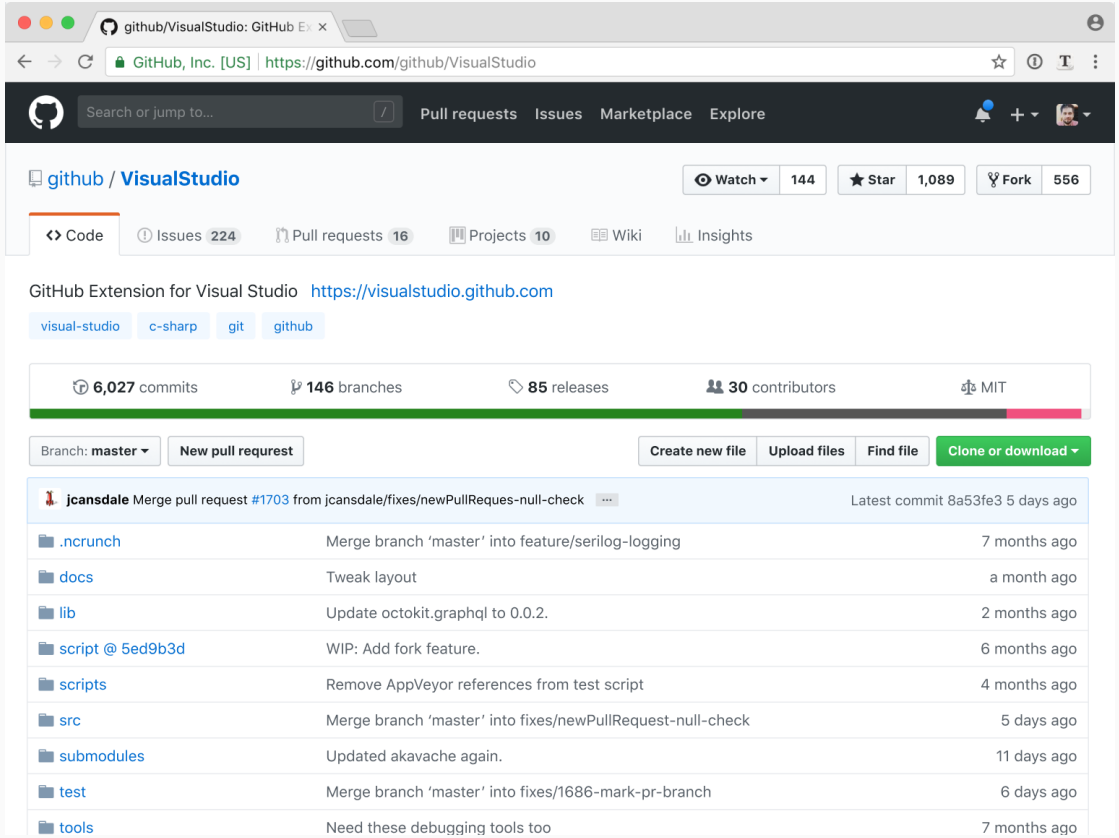
GitHub

GitHub

GitHub is a web-based interface that provides access to open source version control software that lets multiple people make separate changes to projects at the same time.

GitHub Repository

Github is centred around repositories (or repos). A repo is a folder in which all files/folders associated with your project and their version histories are stored.



Working in GitHub

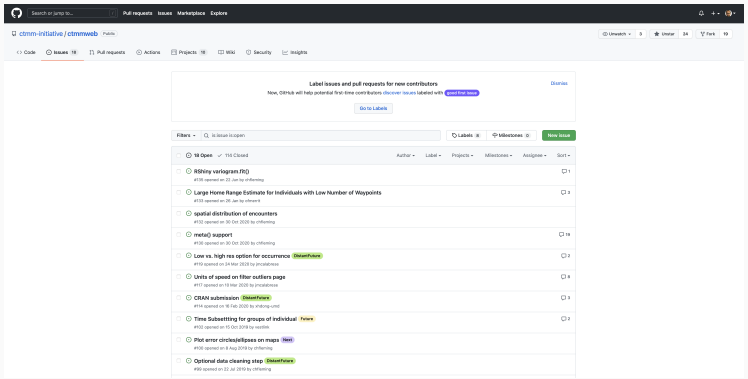
Working in GitHub involves pushing, pulling, merging, and committing changes to branches.

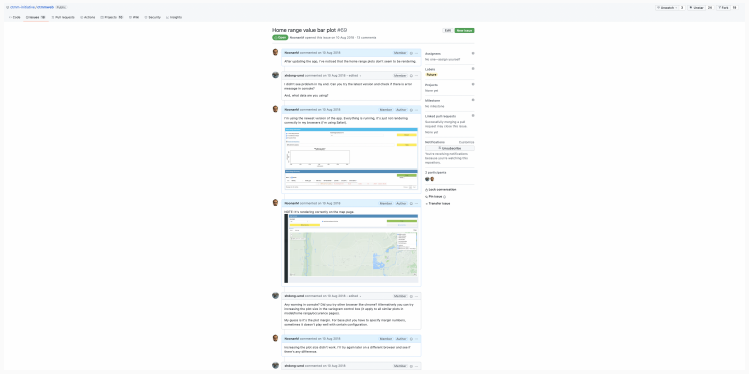
- **Branch** — a workspace in which you can make changes (can be the main branch, or a personal branch).
- **Commit** — a saved record of a change made to a file within the repo.
- **Pull Request (PR)** — the way to ask for changes made to a branch to be merged into another branch that also allows for multiple users to see, discuss and review work being done.
- **Merge** — after a pull request is approved, the commit will be pulled in (or merged) from one branch to another.

Issues

GitHub also provides a place for team members to discuss 'Issues' that need addressing.

In essence the issues feature allow people to identify new tasks that need to be tackled, and to track progress on the task from beginning to end.





GitHub Cons

GitHub repositories can always be changed or deleted (meaning they do not function as permanently stable archives).

GitHub repositories have limited data storage capacity (2Gb), not suitable for some types of research.

GitHub repositories can be public, which can make them inappropriate for sensitive data.

GitHub Summary

GitHub is free, open source, and allows you to house all of your work (data cleaning and analysis, figures, tables, writing) in a single place.

It has built in version control, meaning you can track your changes and recover files (also protects you from computer failures).

A repository can be public or private, giving you full control over how your work is being shared and who can collaborate on your project.

Also allows you to host web pages (e.g., the 520C course website is hosted in GitHub).

Can be integrated into Gitter, permitting a chat interface.

Cluster Computing

Run Times

Most standard analyses are fairly quick and can be run on the order of seconds to minutes.

Large datasets, complex models, bootstrapping, bayesian models, and/or simulation based experiments are becoming increasingly common in biological research.

This can easily push the capacity of standard desktop computers, requiring days to months of computation time.

Cluster Computing

Cluster computing lets you spread your calculations out across multiple, linked computers (nodes).

Leaning on a cluster can dramatically cut down on run times (e.g., from months to days, or days to hours).

As UBC students you have access to Sockeye and Compute Canada (you also have access to Chinook for secure data storage).

If you think you will need to use a cluster as part of your research, UBC's Advanced Research Computing team have a series of webinars on how to use Sockeye: https://osf.io/wpcg6/, and there is a good wiki for Compute Canada.