

Project Management with RStudio and Github

Yujun Zhou

September 21, 2018

Overview

- Why do we need project management ?
- Principles for organized project management
- Introductions to R, RStudio, Rmarkdown, Github
- How to manage project with RStudio and Github
- Readable Code Style

Why do we need project management ?

















 cluster_near_mkt.csv	2/20/2018 12:26	Microsoft Excel C...	38 KB
 cluster_weather.R	2/19/2018 13:25	R File	2 KB
 code.R	5/1/2017 14:37	R File	3 KB
 code_final.R	5/1/2017 14:37	R File	6 KB
 concordance_ipc.dta	1/23/2018 11:28	Stata Dataset	98 KB
 confusion.R	2/14/2018 11:35	R File	24 KB
 daily_rain_2013.dta	12/22/2017 07:52	Stata Dataset	185 KB
 dailyrain.do	2/15/2018 17:39	Stata Do-file	3 KB
 dailyrain.dta	5/5/2017 07:32	Stata Dataset	943 KB
 dailyrain_cluster.dta	2/21/2018 08:56	Stata Dataset	54,256 KB
 December.png	10/31/2017 12:03	PNG File	246 KB
 density.csv	11/3/2017 09:50	Microsoft Excel C...	11,418 KB
 density_plot_code.R	12/12/2017 17:41	R File	1 KB
 density_plots_cluster.docx	2/22/2018 19:17	Microsoft Word D...	963 KB
 Density0.png	11/4/2017 06:36	PNG File	501 KB
 density2.png	11/4/2017 06:34	PNG File	292 KB

Figure 1: Messy project folder

Why do we need project management ?

Raw data, cleaned data, functions,scripts, graphs all in one place

- Original and processed data all in one place: possible contamination of raw data
- Hard to find the exact code to produced a particular table or figure: impossible to reproduce results
- Waste time trying to figure out where to start/proceed after a while

Why do we need project management ?











Name
 Press release for approval.doc
 Press release final.doc
 Press release FINAL VERSION.doc
 Press release FINAL FINAL VERSION.doc
 IMPROVED FINAL PRESS RELEASE.doc
 REVISED APPROVED FINAL PRESS RELEASE.doc
 REVISED APPROVED FINAL PRESS RELEASE v. 2.doc
 !! NEW REVISED APPROVED FINAL PRESS RELEASE v. 2.doc
 !!! REVISED NEW REVISED APPROVED FINAL PRESS RELEASE v. 2.doc
 !!!! Press release as sent.doc

Figure 2: Typical draft versions

Why do we need project management ?

Version control of drafts: Finalfinal_draft_V6.3.docx

- Version names are meaningless, even with dates
- Version notes: run model 0 with county fixed effects, generated figure 1 and table 2; Formatted draft in AJAE style.
- Code and draft are separated: restore results on earlier draft is impossible
- Replicating your work is hard, which makes it hard to collaborate with others

Principles for building an organized project

- ① Build good practices instead of taking the time to clean things up
 - You will NEVER find the time

Principles for building an organized project

- 1 Progress management
 - Start planning as the project lead
 - Define objectives with deadlines: IPAD presentation in two month, paper draft in a year, etc
 - Break down into smaller tasks and then add daily to-do lists
 - Prioritize between projects

Principles for building an organized project

1 Progress management

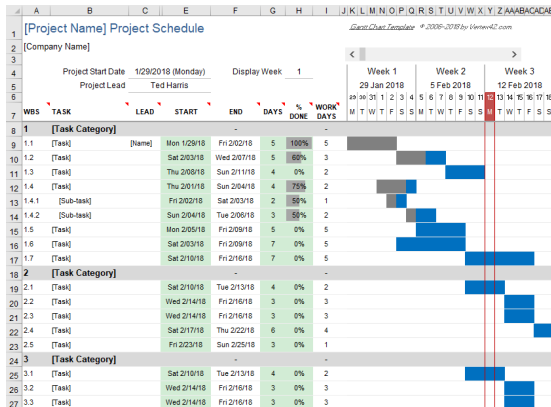
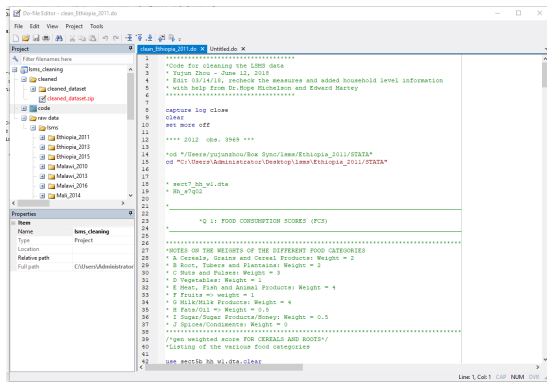


Figure 3: Gantt Chart template in Excel

Principles for building an organized project

2 Use “Project” to manage your project!



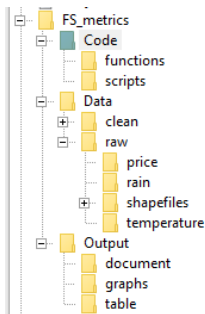
The screenshot shows a Stata project window titled "clean_Ethiopia_2011.do". The interface is divided into three main sections:

- Project Explorer (Left):** Shows a hierarchical view of the project files. The root is "hms_cleaning", which contains subfolders for "cleaned_data", "code", and "raw_data". Under "raw_data", there are folders for "IYDS" and "Ethiopia_2011". The "Ethiopia_2011" folder contains subfolders for the years 2010, 2011, 2012, 2013, 2014, 2015, and 2016.
- Code Editor (Center):** Displays the Stata script "clean_Ethiopia_2011.do". The code includes comments about cleaning DHS data, a date stamp, and commands to set the working directory to the project's STATA folder, load the "wct7_bb_w1.dta" dataset, and generate a weighted score for food consumption scores (FCS) based on various food categories and their weights.
- Properties Panel (Bottom Left):** Shows the properties of the selected "hms_cleaning" project, including its name, type, location, relative path, and full path.

Figure 4: Stata project

Principles for building an organized project

3 Organized layout with relative path



cluster_near_mkt.csv	2/20/2018 12:26	Microsoft Excel C...	38 KB
cluster_weather.R	2/19/2018 13:25	R File	2 KB
code.R	5/1/2017 14:37	R File	3 KB
code_final.R	5/1/2017 14:37	R File	6 KB
concordance_ipc.dta	1/23/2018 11:28	Stata Dataset	98 KB
confusion.R	2/14/2018 11:35	R File	24 KB
daily_rain_2013.dta	12/22/2017 07:52	Stata Dataset	185 KB
dailyrain.do	2/15/2018 17:39	Stata Do-file	3 KB
dailyrain.dta	5/5/2017 07:32	Stata Dataset	943 KB
dailyrain_cluster.dta	2/21/2018 08:56	Stata Dataset	54,256 KB
December.png	10/31/2017 12:03	PNG File	246 KB
density.csv	11/3/2017 09:50	Microsoft Excel C...	11,418 KB
density_plot_code.R	12/12/2017 17:41	R File	1 KB
density_plots_cluster.docx	2/22/2018 19:17	Microsoft Word D...	963 KB
Density0.png	11/4/2017 06:36	PNG File	501 KB
density2.png	11/4/2017 06:34	PNG File	292 KB

Manage files with relative path

- Separate raw and processed data: raw data should never be touched
`write.csv(choma_maize_price, "data/clean/choma_maize.csv")`
- Separate cleaning and analysis scripts
- Separate inputs and outputs (drafts, graphs, tables)
`ggsave("output/graphs/myplot.png")`

Use Version Control Tools

- Time capsule for entire projects

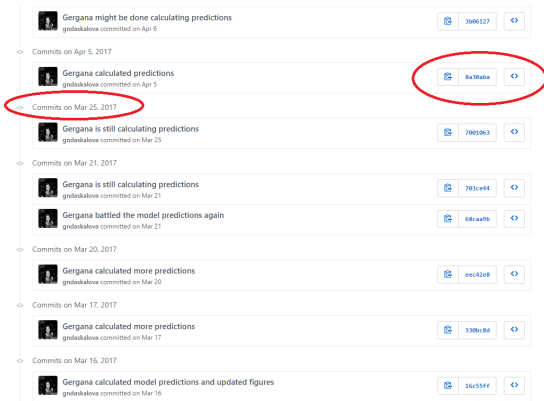


Figure 5: Version history

R, RStudio, Rmarkdown, Github

- R: Open source statistical computing software, flexible data structure, multiple purpose.
- RStudio: an open source development environment for R with nice GUI. Your hub to everything else: Github, latex, website management etc.
- Markdown/ RMarkdown: statistical analysis and results contained in documents
- Easiest way to formulate math formulas (Bye ! Latex compiling errors)
- Formats: PDF, Word, HTML, Beamer
- Github is a cloud-based repository for version control. You can store your code, collaborate on someone else's project, etc.

Benefits of using RStudio and Github integration

- A Github repository = R project
- Easy to work on multiple device
- Click to commit and push


Getting Started

- Sign up for Github and set up your profile
- Create a repository

Create a new repository


A repository contains all the files for your project, including the revision history.


Owner Repository name

 gndaskalova ▾ / ✓

Great repository names are short and memorable. Need inspiration? How about [improved-umbrella](#).

Description (optional)

 **Public**
Anyone can see this repository. You choose who can commit.

 **Private**
You choose who can see and commit to this repository.

Initialize this repository with a README
This will let you immediately clone the repository to your computer. Skip this step if you're importing an existing repository.

▾ ▾ ⓘ

Figure 6: Create repository

Link Git in RStudio

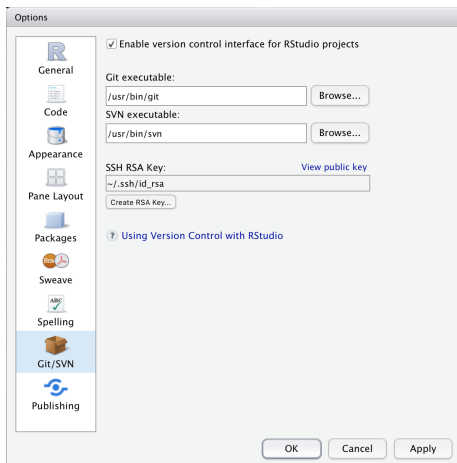


Figure 7: Git setup

How to manage Project with RStudio

- Create a local project from your Github repository

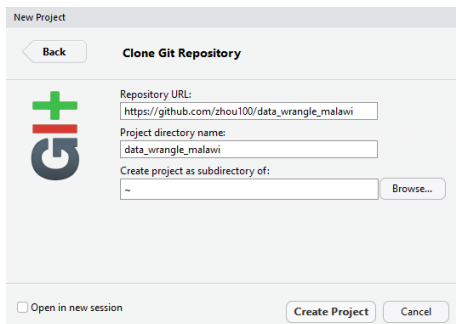


Figure 8: Create project from Git repository

How to manage Project with RStudio

- Commit local changes to online repository, then click push

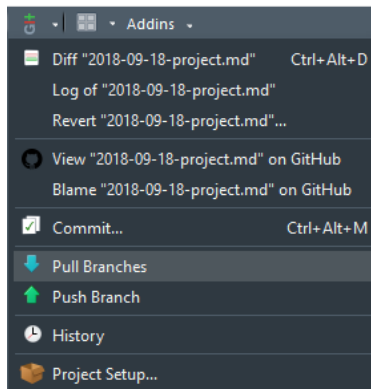


Figure 9: Commit, Pull and Push

Summary

- Make your life easier using projects with version control
- Make collaborations easier with lab-mates, coauthors and adviser
- Prof. Mindy Mallory's "Reproducible Research Practices for Economists"
- Demonstrate End-to-End project management ability
- Graduate school is project-based learning experience

Github as a technical CV

The Octocat
octocat

Unfollow

Block or report user

Developer Program Member

GitHub STAFF

San Francisco

octocat@github.com

https://blog.github.com/

Organizations

Overview Repositories 205 Stars 920 Followers 6.3k Following 538

Pinned repositories

- jekyl/jekyll**
Jekyll is a blog-aware, static site generator in Ruby.
Ruby ★ 35k ¥ 2.7k
- wordpress-to-jekyll-exporter**
One-click WordPress plugin that converts all posts, pages, taxonomies, metadata, and settings to Markdown and YAML, which can be dropped into Jekyll.
PHP ★ 805 ¥ 105
- github/choosealicense.com**
A site to provide non-judgmental guidance on choosing a license for your open source project.
Ruby ★ 1.3k ¥ 392
- word-to-markdown**
A ruby gem to liberate content from Microsoft Word documents.
Ruby ★ 945 ¥ 102
- licensee**
A Ruby Gem to detect under what license a project is distributed.
Ruby ★ 265 ¥ 83
- jekyl/jekyll-admin**
A Jekyll plugin that provides users with a traditional CMS-style graphical interface to author content and administer Jekyll sites.
JavaScript ★ 1.7k ¥ 193

4,675 contributions in the last year

2018

2017

2016

2015

2014

2013

2012

2011

2010

2009

2008

2007

2006

...

Activity overview

Contributed to github/github, github/pages, and 5 other repositories

Activity overview

27% Code review

47% Commits

14% Issues

10% Pull requests

Contribution activity

Jump to -

August 2018

Created 23 commits in 5 repositories

- Separate Functions from Scripts
 - Functions are for specific purposes and can be used elsewhere
 - Scripts are used to generate graphs and tables and save them as intermediate products
 - Make the body of the scripts more readable

Script vs Function

```
1  # Goal : This script aims to clean up wfp price data, impute missing price
2  # purpose: use the clean market price and thinness measure to generate the
3
4
5  # Input :
6  # 1. raw csv files downloaded from wfp price
7  # 2. population density raster data
8  # 3. coordinates of clusters.
9  # 4. shapefile of livelihood zones.
10
11 # Output:
12 # 0. market coordinates generated from market_coordinates.R
13 # 1. a df of price by product by mkt by yearmon, with missing imputed by n
14 # 2. a df of market thinness measures by product by mkt by yearmon
15 # 3. matching the price and mkt_thinness to the cluster and ipzone level
16 #
17 # Yujun Zhou - 03/20/18
18 #####
19
20
21 package = c("dplyr","maptools","rgeos", "rgdal", "raster")
22 lapply(package, require, character.only = TRUE)
23
24 source("R/functions/Yearmon.R")
25 source("R/functions/market_transpose.R")
26 source("R/functions/NearMkt.R")
27 source("R/functions/spatial_price_impute.R")
28 source("R/functions/Popuweight.R")
29 source("R/functions/NameToPrice.R")
30 source("R/functions/WeightedPrice.R")
31 source("R/functions/MktReshape.R")
32
33 #####
```

Figure 11: Scripts

Script vs Function

```
#####  
# Goal : generate Date and yearmon variable to help with join  
# input: data frame with year and month variables  
# output: data frame with Date and yearmon  
#####  
library(zoo)  
  
yearmon = function(df,year_var,month_var){  
  
  # Find the year and month stored in the dataframe  
  year = df[[year_var]]  
  month = df[[month_var]]  
  
  # Transform month to character  
  month_character<-month.abb[month]  
  
  # Year-month in character  
  yearmon_character = paste(month_character,year,sep="/")  
  
  # year-month in numbers  
  yearmon_format<-as.yearmon(yearmon_character,format = "%b/%Y")  
  
  # Date format  
  date_format<-as.Date(yearmon_format)  
  
  # save yearmon and date in a dataframe |  
  df[["yearmon"]] = yearmon_format  
  df[["date"]] = date_format  
  
  return(df)  
}
```

Figure 12: Function

Readable Code Style

- Comment at the top about the purpose of the code, input and output
- Use packages and source functions at the start
- Make comments on steps
- Use meaningful variable or function names
- More on code style: [Google's R Style Guide](#)

Readable Code Style

- Use dplyr package and piping
 - Example: list the median size of each type (at least 3), in decreasing order

```
data %>% group_by(type) %>%  
  
  summarise(median_size = median(size, na.rm = TRUE)) %>%  
  
  filter(median_size > 3) %>%  
  
  arrange(desc(median_size)) %>%  
  
  select(type, median_size)
```