Reproducible Coding

Best Practices

Jennifer Spillane - Lab Meeting - February 27

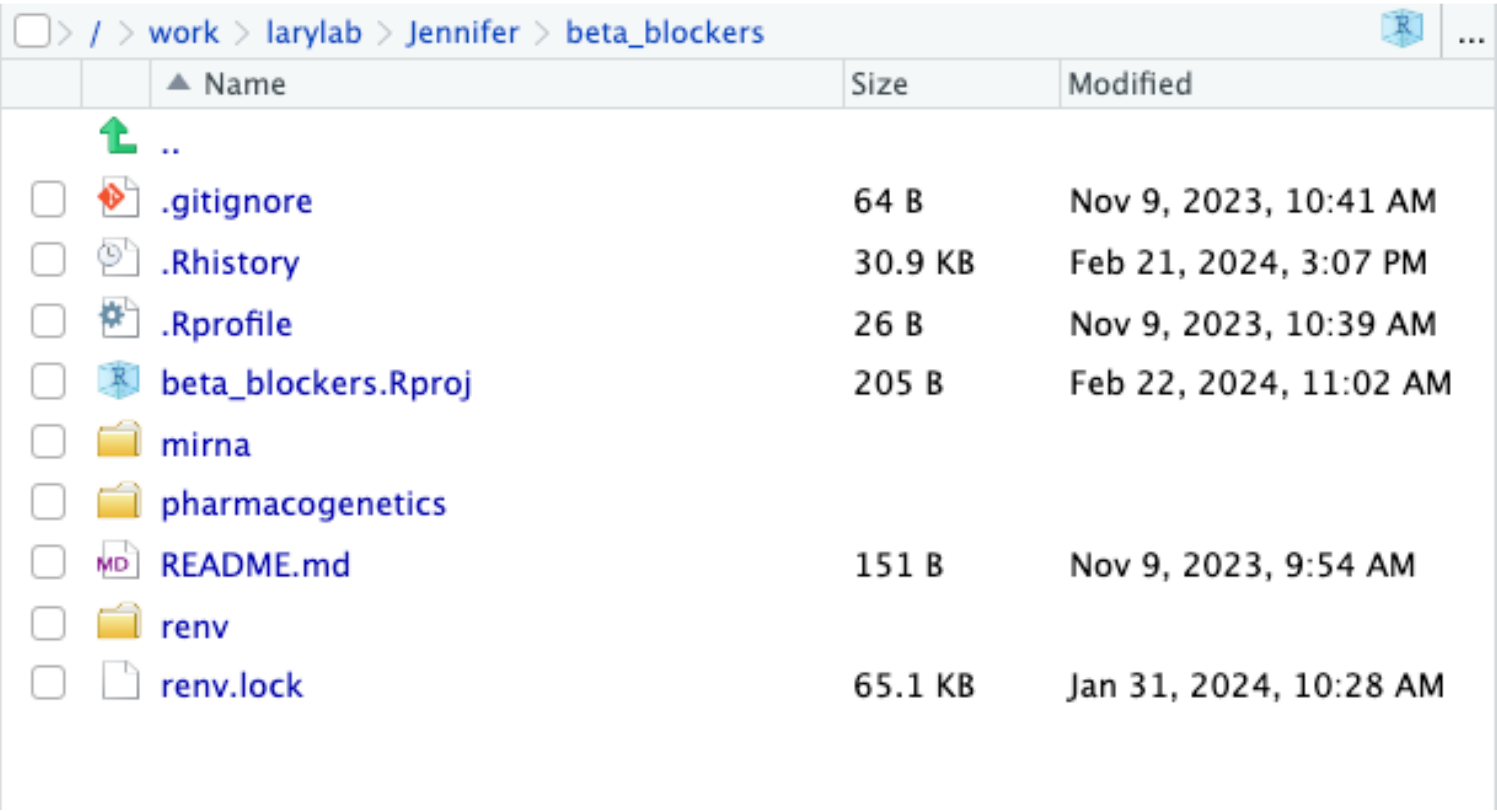
(Push your code to GitHub)

Project Organization

- In most projects (especially ones that you are in charge of or are working on alone):
    - **1 main project directory** - should hold everything you create for the project
        - This does *not* include the actual original data - good to keep this in a dedicated data location (/work/larylab/data/name_of_your_project/)
    - 3 directories inside the project directory
        - **scripts** - to hold all scripts used to analyze and plot data (including slurm scripts)
        - **data** - to hold intermediate data files that you generate over your analysis
        - **results** - to hold all results files, especially tables and figures
        - Any of these can have sub-directories if it's helpful/keeps things more organized

(Push your code to GitHub)

Beta Blocker Example

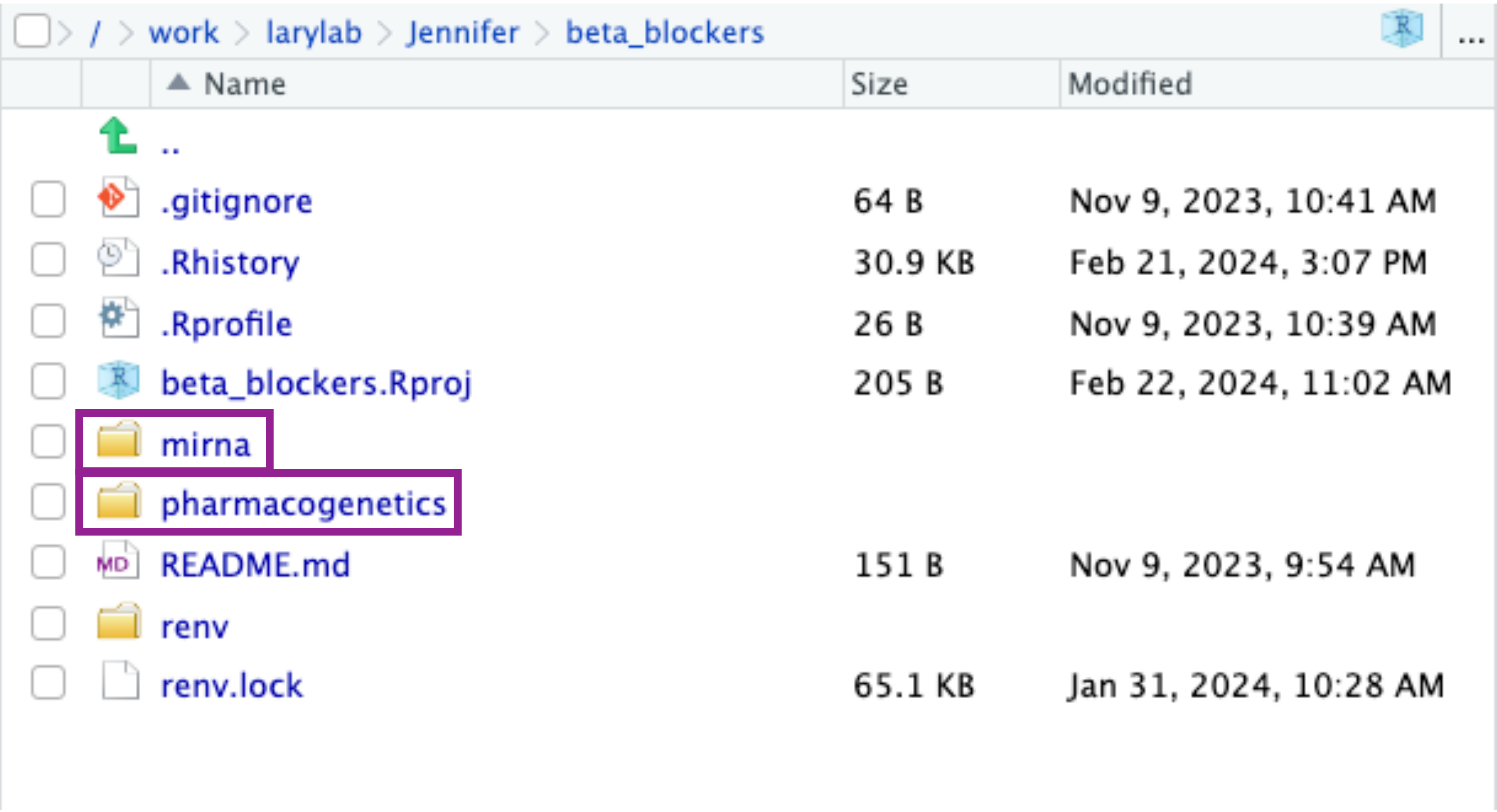


/ > work > larylab > Jennifer > beta_blockers

	▲ Name	Size	Modified
	..		
□	.gitignore	64 B	Nov 9, 2023, 10:41 AM
□	.Rhistory	30.9 KB	Feb 21, 2024, 3:07 PM
□	.Rprofile	26 B	Nov 9, 2023, 10:39 AM
□	beta_blockers.Rproj	205 B	Feb 22, 2024, 11:02 AM
□	mirna		
□	pharmacogenetics		
□	README.md	151 B	Nov 9, 2023, 9:54 AM
□	renv		
□	renv.lock	65.1 KB	Jan 31, 2024, 10:28 AM

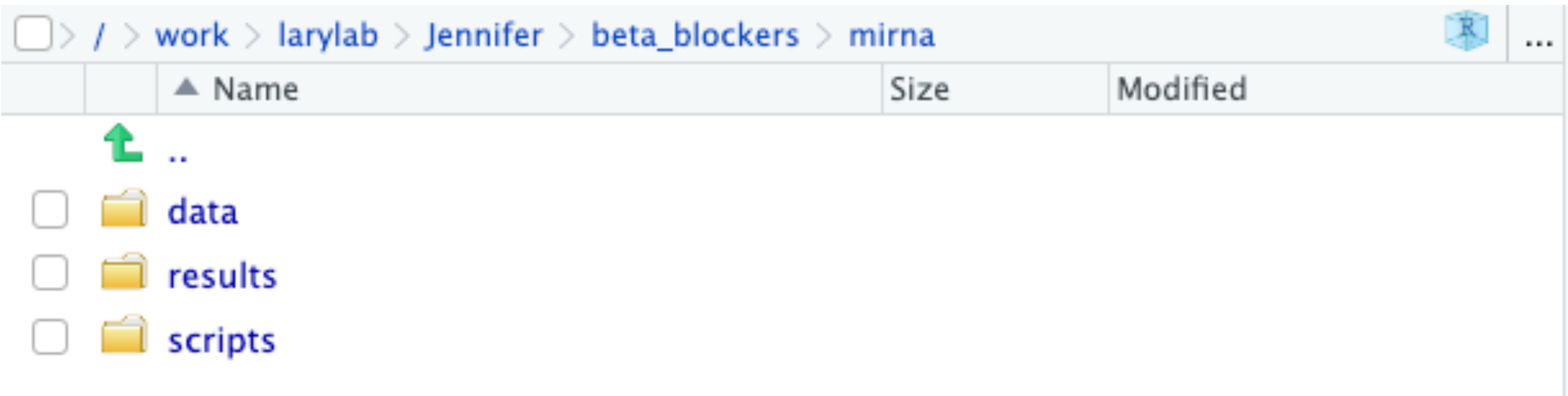
(Push your code to GitHub)

Beta Blocker Example



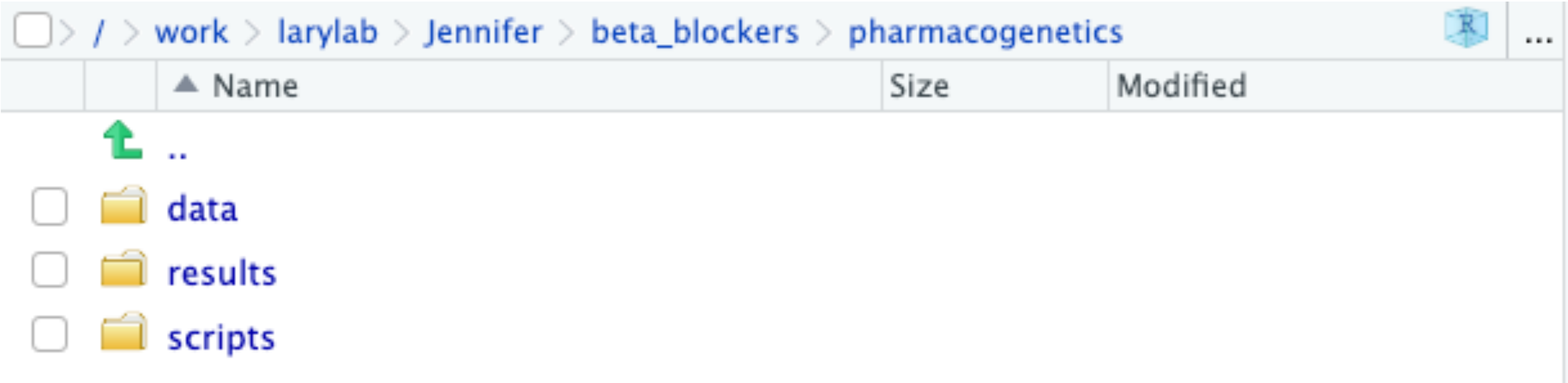
/ > work > larylab > Jennifer > beta_blockers

	Name	Size	Modified
	..		
□	.gitignore	64 B	Nov 9, 2023, 10:41 AM
□	.Rhistory	30.9 KB	Feb 21, 2024, 3:07 PM
□	.Rprofile	26 B	Nov 9, 2023, 10:39 AM
□	beta_blockers.Rproj	205 B	Feb 22, 2024, 11:02 AM
□	mirna		
□	pharmacogenetics		
□	README.md	151 B	Nov 9, 2023, 9:54 AM
□	renv		
□	renv.lock	65.1 KB	Jan 31, 2024, 10:28 AM



/ > work > larylab > Jennifer > beta_blockers > mirna

	Name	Size	Modified
	..		
□	data		
□	results		
□	scripts		



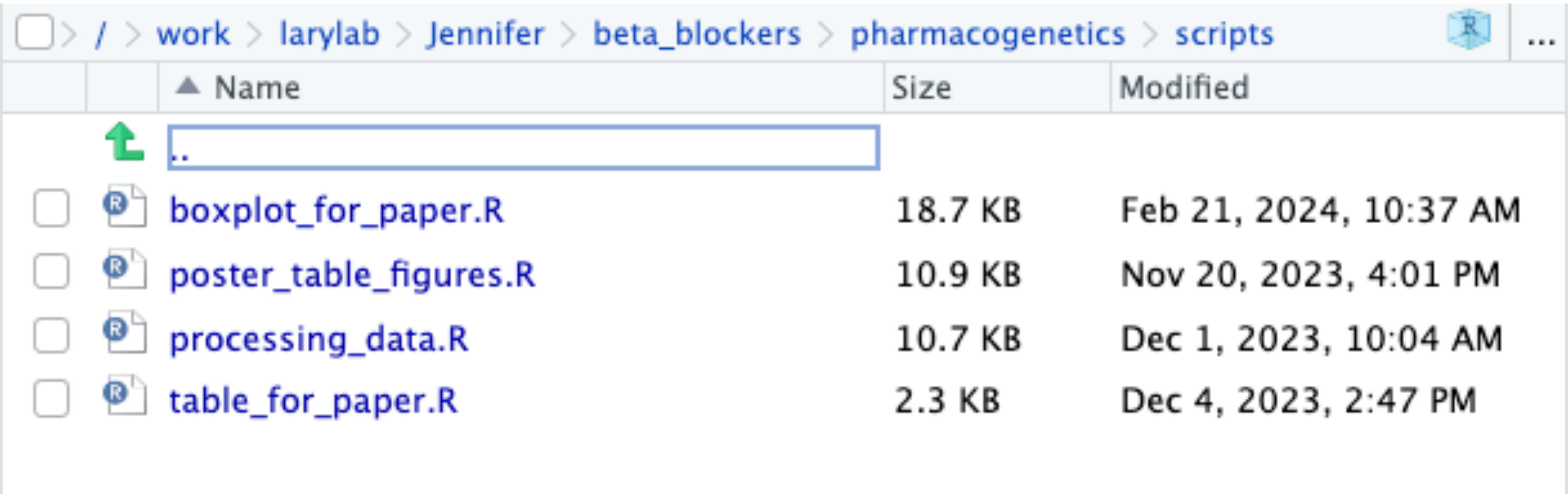
/ > work > larylab > Jennifer > beta_blockers > pharmacogenetics

	Name	Size	Modified
	..		
□	data		
□	results		
□	scripts		

(Push your code to GitHub)

“scripts” Directory

- Should have all the scripts needed to replicate your analysis
    - 1 script for generating an “analytic file” - the file you’ll primarily use for making tables, plotting, fitting models, etc.
    - Other scripts for actually making the tables and doing the stats
    - Include slurm scripts for each process, but representative examples are okay



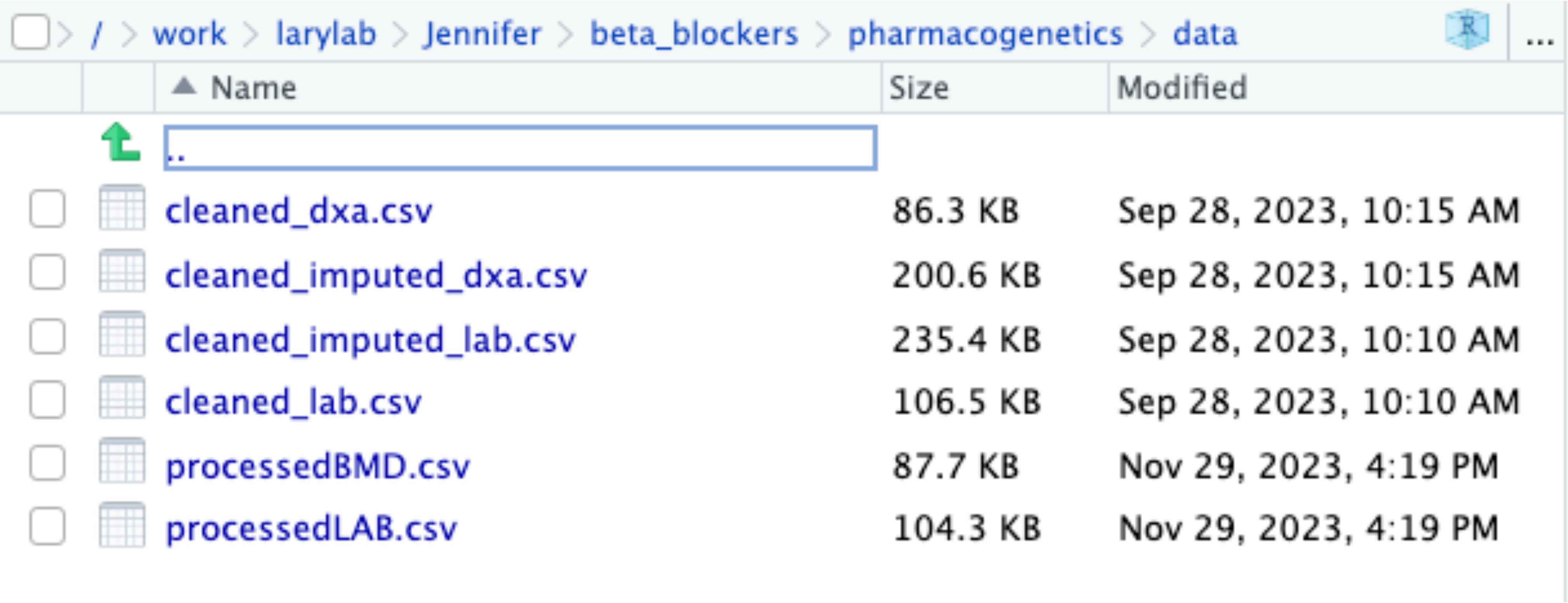
/ > work > larylab > Jennifer > beta_blockers > pharmacogenetics > scripts

▲ Name	Size	Modified
..		
boxplot_for_paper.R	18.7 KB	Feb 21, 2024, 10:37 AM
poster_table_figures.R	10.9 KB	Nov 20, 2023, 4:01 PM
processing_data.R	10.7 KB	Dec 1, 2023, 10:04 AM
table_for_paper.R	2.3 KB	Dec 4, 2023, 2:47 PM

(Push your code to GitHub)

“data” Directory

- Should have all the intermediate files you generate over the course of your analysis
  - It’s usually a lot of steps to get from raw data files to finished analyses
  - Writing out intermediate files to your data directory ensures that you don’t have to repeat the data processing steps tons of times when all you want to do is change a figure
  - Can write out to a csv file if you might need to port to the command line or a different language, or an RData file if the whole analysis will take place in R.



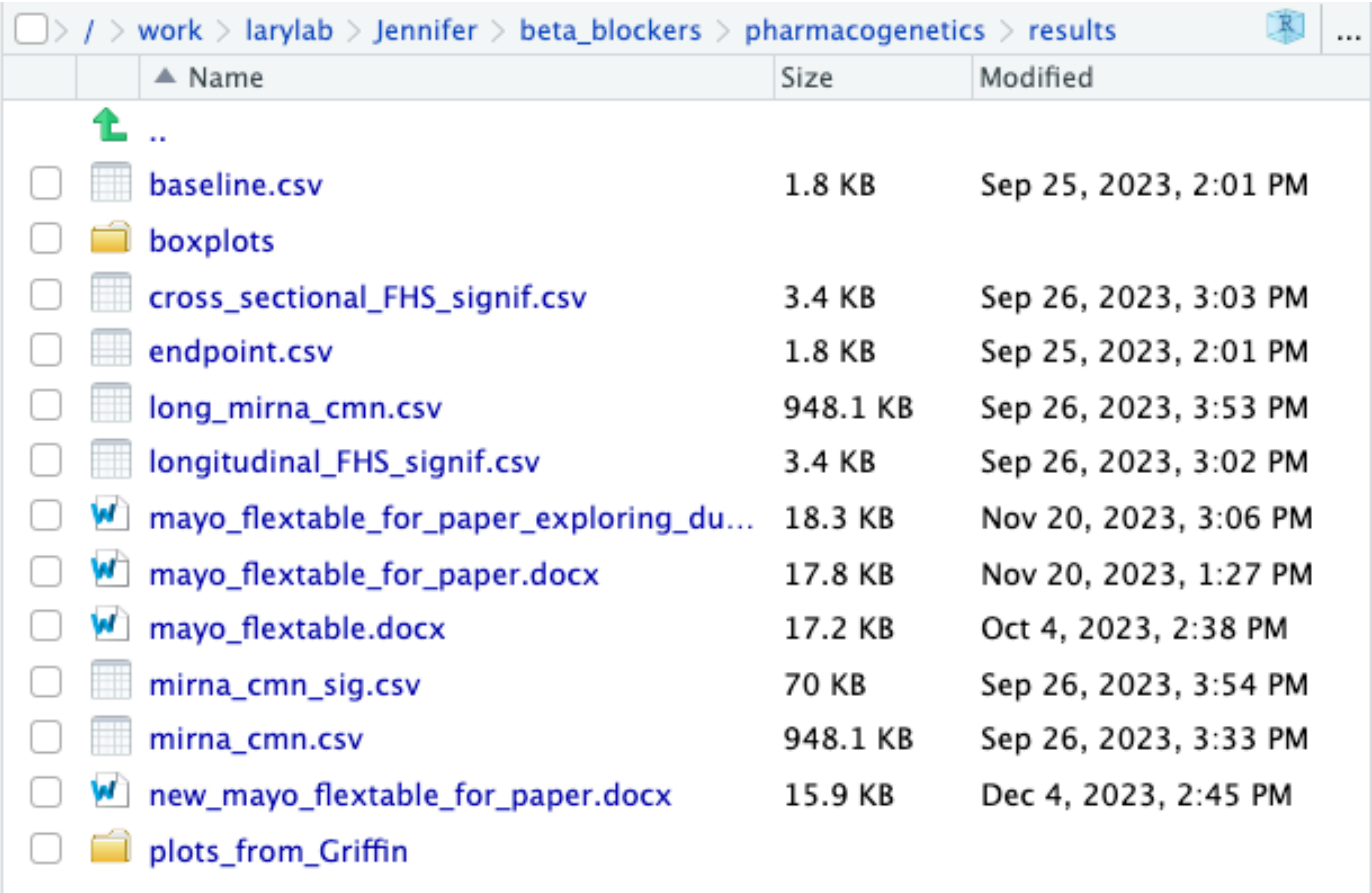
/ > work > larylab > Jennifer > beta_blockers > pharmacogenetics > data

Name	Size	Modified
..		
cleaned_dxa.csv	86.3 KB	Sep 28, 2023, 10:15 AM
cleaned_imputed_dxa.csv	200.6 KB	Sep 28, 2023, 10:15 AM
cleaned_imputed_lab.csv	235.4 KB	Sep 28, 2023, 10:10 AM
cleaned_lab.csv	106.5 KB	Sep 28, 2023, 10:10 AM
processedBMD.csv	87.7 KB	Nov 29, 2023, 4:19 PM
processedLAB.csv	104.3 KB	Nov 29, 2023, 4:19 PM

(Push your code to GitHub)

“results” Directory

- This is where I often have the most sub-directories
- How can you tell if a file should go here or in the data directory?
  - If I use that file in another script, I put it in the data directory, otherwise here



/ > work > larylab > Jennifer > beta_blockers > pharmacogenetics > results

	Name	Size	Modified
	..		
□	baseline.csv	1.8 KB	Sep 25, 2023, 2:01 PM
□	boxplots		
□	cross_sectional_FHS_signif.csv	3.4 KB	Sep 26, 2023, 3:03 PM
□	endpoint.csv	1.8 KB	Sep 25, 2023, 2:01 PM
□	long_mirna_cmn.csv	948.1 KB	Sep 26, 2023, 3:53 PM
□	longitudinal_FHS_signif.csv	3.4 KB	Sep 26, 2023, 3:02 PM
□	mayo_flextable_for_paper_exploring_du...	18.3 KB	Nov 20, 2023, 3:06 PM
□	mayo_flextable_for_paper.docx	17.8 KB	Nov 20, 2023, 1:27 PM
□	mayo_flextable.docx	17.2 KB	Oct 4, 2023, 2:38 PM
□	mirna_cmn_sig.csv	70 KB	Sep 26, 2023, 3:54 PM
□	mirna_cmn.csv	948.1 KB	Sep 26, 2023, 3:33 PM
□	new_mayo_flextable_for_paper.docx	15.9 KB	Dec 4, 2023, 2:45 PM
□	plots_from_Griffin		

(Push your code to GitHub)

Scripting Best Practices

What your scripts look like while working on a project will be different than how they look at the end!

- Scripts should:
    - Have logical names
    - Include author information
    - Have info about what the script is doing/goals
    - Load packages and establish a working directory at the beginning
    - Read in data files at the top
    - Have consistent spacing throughout (both for indentations and around punctuation)

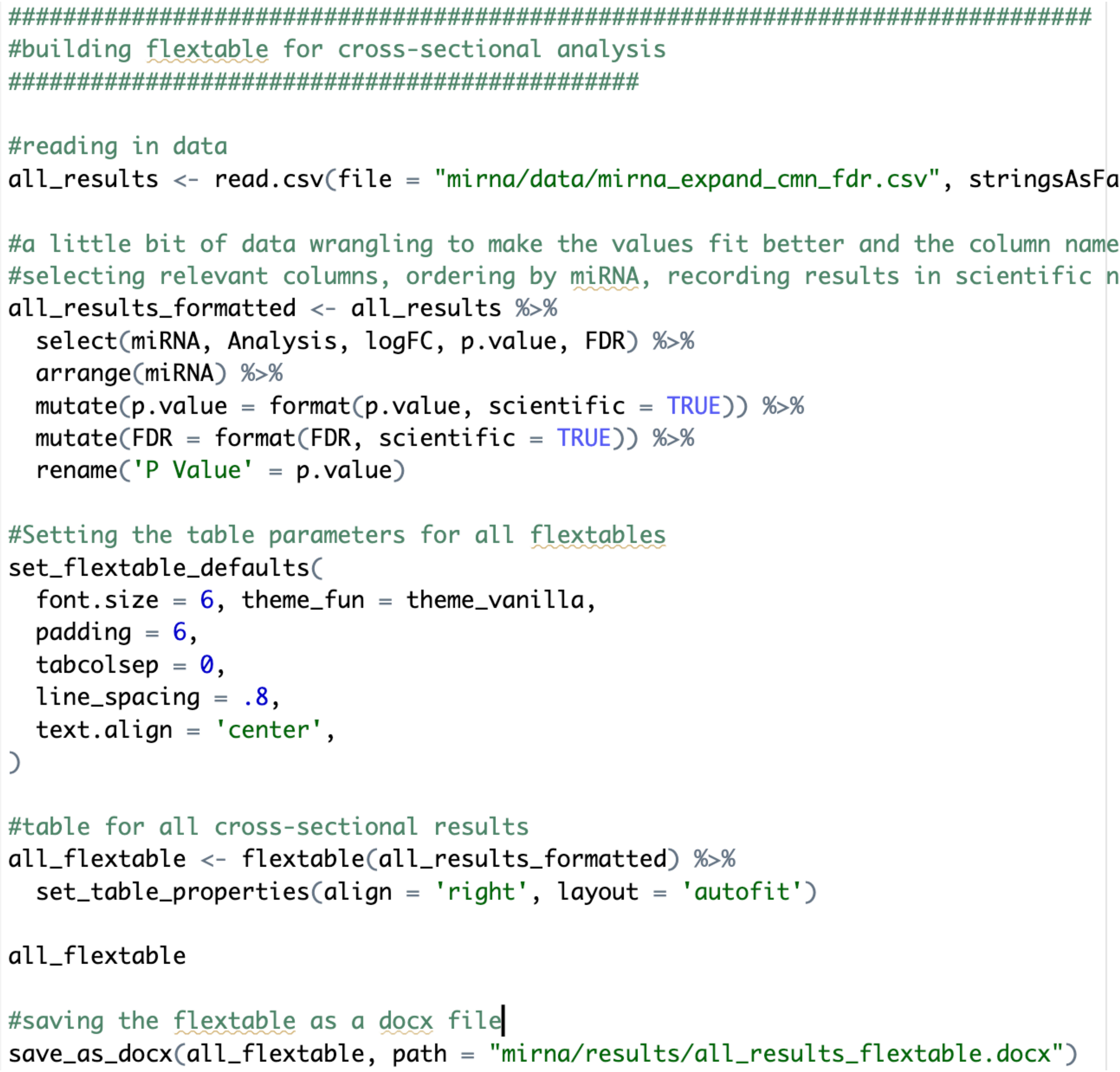
(Push your code to GitHub)

Scripting Best Practices

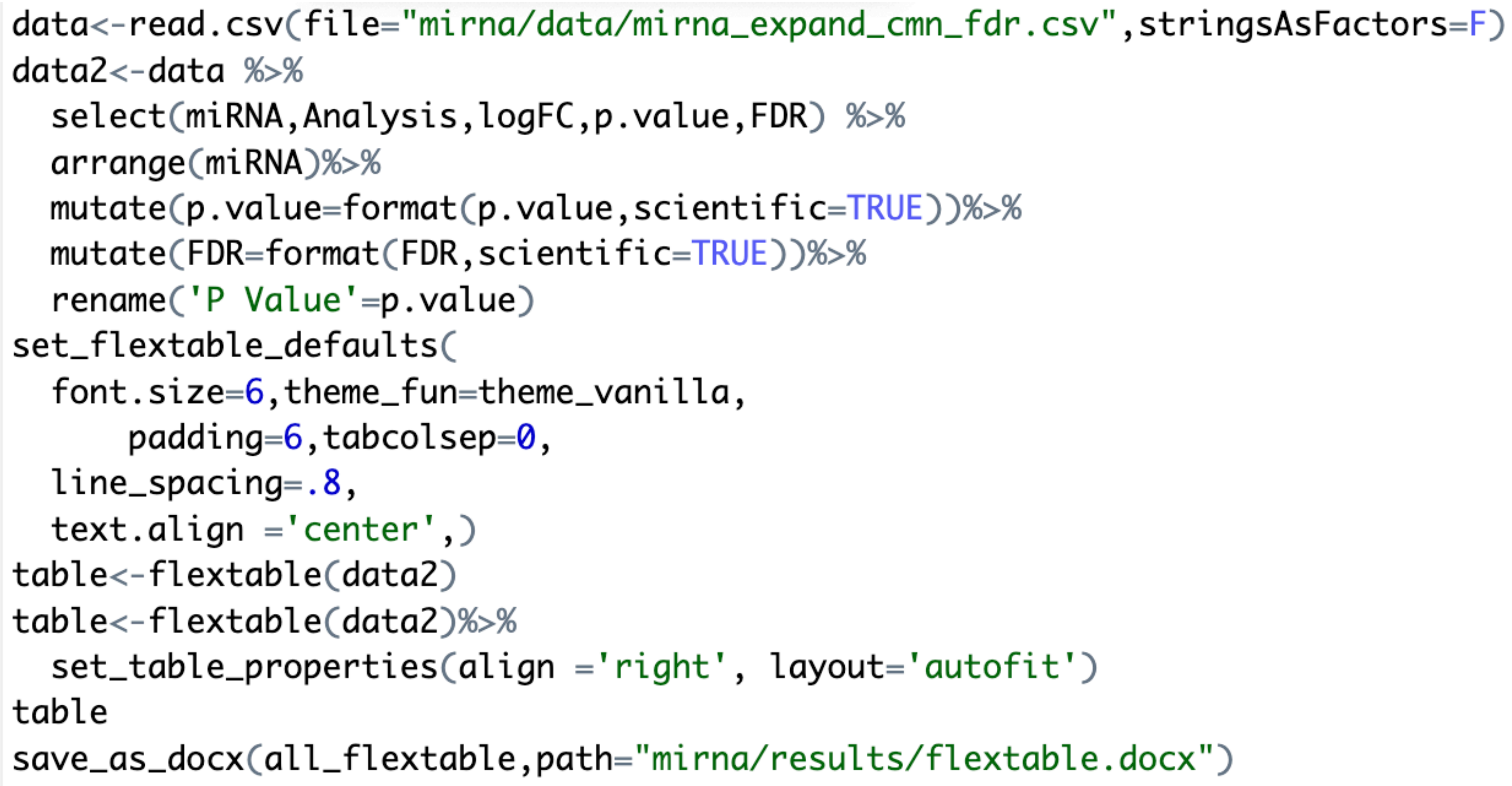
What your scripts look like while working on a project will be different than how they look at the end!

- Scripts should:
  - Be commented every few lines/whenever you begin a new process
  - Have object/variable names that are informative
  - Not contain code that doesn't work/isn't necessary anymore
    - If you aren't sure, comment it out, clear your environment and then run the code again
  - Logically (and clearly) feed into one another
  - Be able to be run top to bottom with no errors

(Push your code to GitHub)



```
###############################################################################
#building flextable for cross-sectional analysis
##############################################

#reading in data
all_results <- read.csv(file = "mirna/data/mirna_expand_cmn_fdr.csv", stringsAsFactors = F)

#a little bit of data wrangling to make the values fit better and the column names more readable
#selecting relevant columns, ordering by miRNA, recording results in scientific notation, and renaming columns
all_results_formatted <- all_results %>%
  select(miRNA, Analysis, logFC, p.value, FDR) %>%
  arrange(miRNA) %>%
  mutate(p.value = format(p.value, scientific = TRUE)) %>%
  mutate(FDR = format(FDR, scientific = TRUE)) %>%
  rename('P Value' = p.value)

#Setting the table parameters for all flextables
set_flextable_defaults(
  font.size = 6, theme_fun = theme_vanilla,
  padding = 6,
  tabcolsep = 0,
  line_spacing = .8,
  text.align = 'center',
)

#table for all cross-sectional results
all_flextable <- flextable(all_results_formatted) %>%
  set_table_properties(align = 'right', layout = 'autofit')

all_flextable

#saving the flextable as a docx file
save_as_docx(all_flextable, path = "mirna/results/all_results_flextable.docx")
```



```
data<-read.csv(file="mirna/data/mirna_expand_cmn_fdr.csv",stringsAsFactors=F)
data2<-data %>%
  select(miRNA,Analysis,logFC,p.value,FDR) %>%
  arrange(miRNA)%>%
  mutate(p.value=format(p.value,scientific=TRUE))%>%
  mutate(FDR=format(FDR,scientific=TRUE))%>%
  rename('P Value'=p.value)
set_flextable_defaults(
  font.size=6,theme_fun=theme_vanilla,
      padding=6,tabcolsep=0,
  line_spacing=.8,
  text.align ='center',)
table<-flextable(data2)
table<-flextable(data2)%>%
  set_table_properties(align ='right', layout='autofit')
table
save_as_docx(all_flextable,path="mirna/results/flextable.docx")
```

(Push your code to GitHub)

Scripting Best Practices

What your scripts look like while working on a project will be different than how they look at the end!

- Paths in scripts:
    - For *original* data files, include the full path, since these will be stored outside the project directory
    - Set your working directory to the project directory, and then make all other paths *relative* so that they all branch appropriately from the project directory
        - Helps to eliminate accidental duplications of directories
        - Makes the script work better with git version control and collaboration
    - Don't change your working directory after setting it at the top

(Push your code to GitHub)

Scripting Best Practices

What your scripts look like while working on a project will be different than how they look at the end!

- Extra files in support of scripts:
    - Markdown files - can be helpful for writing instructions, recording reasoning behind analytical choices, or documenting code from the command line
    - README files:
        - Beginning of project: should be generally informative - short is fine
        - End of project (close to paper submission): **list the scripts** in the order they should be run, **note the files they use** and the **files they generate** - should be a **guide to reproduce the analysis** (in case someone else wants to use your method or a reviewer wants to check results)

(Push your code to GitHub)