Code Library

Ruhika Chatterjee

2024-08-16

Intro Information

Every Data Science Project starts with a question that is to be answered with data

- That means that forming the question is an important first step in the process.
- The second step is finding or generating the data you're going to use.
- With the question solidified and data in hand, the data are then analyzed first by exploring the data and then often by modeling the data.
- After drawing conclusions, the project has to be communicated to others.
- Most projects build off someone else's work. It's really important to give credit

R project: Data, Scripts, Output

Troubleshooting

- Check error messages and outputs
- Talk to your rubber duck
- https://www.r-project.org/help.html
- Use help() function and ? command (i.e. help(lm) or help("lm"), or ?lm or ¿'lm")
- Use Google or search.r-project.org
- Check forums StackOverflow and CrossValidated

How to effectively ask questions on forums

- The question you are trying to answer
- How you approached the problem, what steps you have already taken
- What steps will reproduce the problem (including sample data!)
- What was the expected output
- What you saw instead (including any error messages you received!)
- What troubleshooting steps you have already tried
- Details about your set-up, eg: OS, Rversion, packages
- Be specific in the title of your questions!
- Read the forum posting guidelines, ask your question on an appropriate forum!
- Be explicit, detailed, courteous, and succinct
- Follow up on the post OR post the solution

Intro to R

- CRAN: https://cran.r-project.org/

- RStudio: https://posit.co/products/open-source/rstudio/
- File Types: RScript, RMarkdown
- R Project (folder/directory for organization): Create directories: Data, Scripts, Output

R Markdown

- https://rmarkdown.rstudio.com/
- **bold** and *italics*
- Parameter {r Markdown, echo = False} to not print code
- Ctrl+Alt+I (Windows) to initiate code block

```
#install.packages("rmarkdown")
```

Version Control

- GitHub Repository Instruction: https://docs.github.com/en/get-started/start-your-journey/hello-world
- Git Download: https://git-scm.com/downloads

To configure Git in Git Bash

- git config –global user.name "Jane Doe"
- git config –global user.email janedoe@gmail.com
- git config –list confirm changes
- exit exit Terminal

Link

- Link RStudio to Git: Tools > Global Options > Git/SVN then confirm directory
- Link to GitHub: "Create RSA Key", close, "View public key", in GitHub Account Settings > "SSH and GPG keys" > "New SSH key", paste, title

Linking Repositories

- To Link Directly: Create repository in GitHub, copy URL, create new project in RStudio (Version control), Git as VC software, paste URL
- Stage, commit, push (with a message explaining what changed, why and by whom) Save, Git Tab check file as "Staged", click Commit, Commit message, Commit, Push
- Clone Repository: RStudio File > New Project > Version Control, Git, URL, location, create

Add existing Project to GitHub

in Git Bash

- cd ~/dir/name/of/path/to/file #navigate Terminal
- git init #initializes as git repo
- git add . #adds directory files to repo
- git commit -m "Initial commit" #commit

in GitHub: create repo (same name, no .readme, .gitignore, license), select "Push an existing repository from the command line", copy the code, reopen all

Basic R Functions

R Data Structures

```r
# Matrix
x <- matrix(c(1, 2, 3, 4, 5, 6, 7, 8), nrow = 4, ncol = 2) #create matrix
```

R Packages

- Repositories: CRAN, BioConductor (bioinformatics), GitHub
- Search: https://www.rdocumentation.org/

```r
version #R info version
```

```
##               _
## platform      x86_64-w64-mingw32
## arch          x86_64
## os            mingw32
## crt           ucrt
## system        x86_64, mingw32
## status
## major         4
## minor         4.1
## year          2024
## month         06
## day           14
## svn rev       86737
## language      R
## version.string R version 4.4.1 (2024-06-14 ucrt)
## nickname      Race for Your Life
```

```r
sessionInfo() #R info version, packages
```

```
## R version 4.4.1 (2024-06-14 ucrt)
## Platform: x86_64-w64-mingw32/x64
## Running under: Windows 11 x64 (build 22631)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
```

```
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.4.1    fastmap_1.2.0     cli_3.6.3         tools_4.4.1
##  [5] htmltools_0.5.8.1 rstudioapi_0.17.1 yaml_2.3.10       rmarkdown_2.29
##  [9] knitr_1.49        xfun_0.49         digest_0.6.37     rlang_1.1.4
## [13] evaluate_1.0.1
```

```r
# Install from CRAN:
#   install.packages("ggplot2", repos = "http://cran.us.r-project.org") #install
#   install.packages(c("labeling","tibble"), repos = "http://cran.us.r-project.org") #multiple

# Install from Bioconductor
#   install.packages("BiocManager", repos = "https://bioconductor.org/biocLite.R")
#   BiocManager::install(c("GenomicFeatures", "AnnotationDbi")) #install package

# Install from GitHub (need package, author name)
#   install.packages("devtools", repos = "http://cran.us.r-project.org") #only once
#   library(devtools)
#   install_github("author/package") #installs package

# library(ggplot2)# Load package, careful of dependencies
# installed.packages() #check installed packages
# library() #alternate
# old.packages(repos = "http://cran.us.r-project.org") #check packages to update
# update.packages(repos = "http://cran.us.r-project.org") #update all packages
# install.packages("ggplot2") #to update single package
# detach("package:ggplot2", unload=TRUE) #unload function
# remove.packages("ggtree") #remove package
# help(package = "ggplot2") #package info
# browseVignettes("ggplot2") #extended help files
```