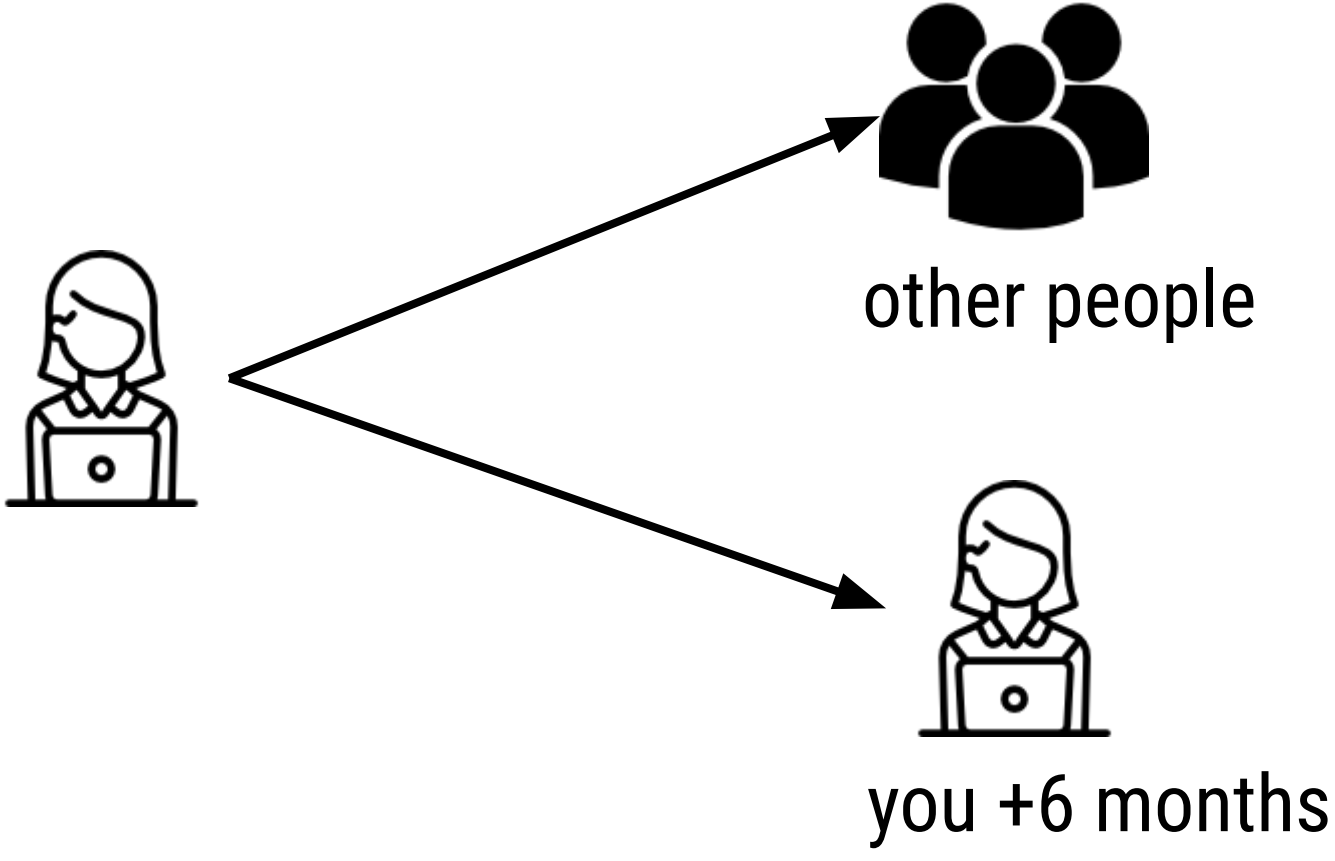
Organizing your data analysis

Getting and Cleaning Data

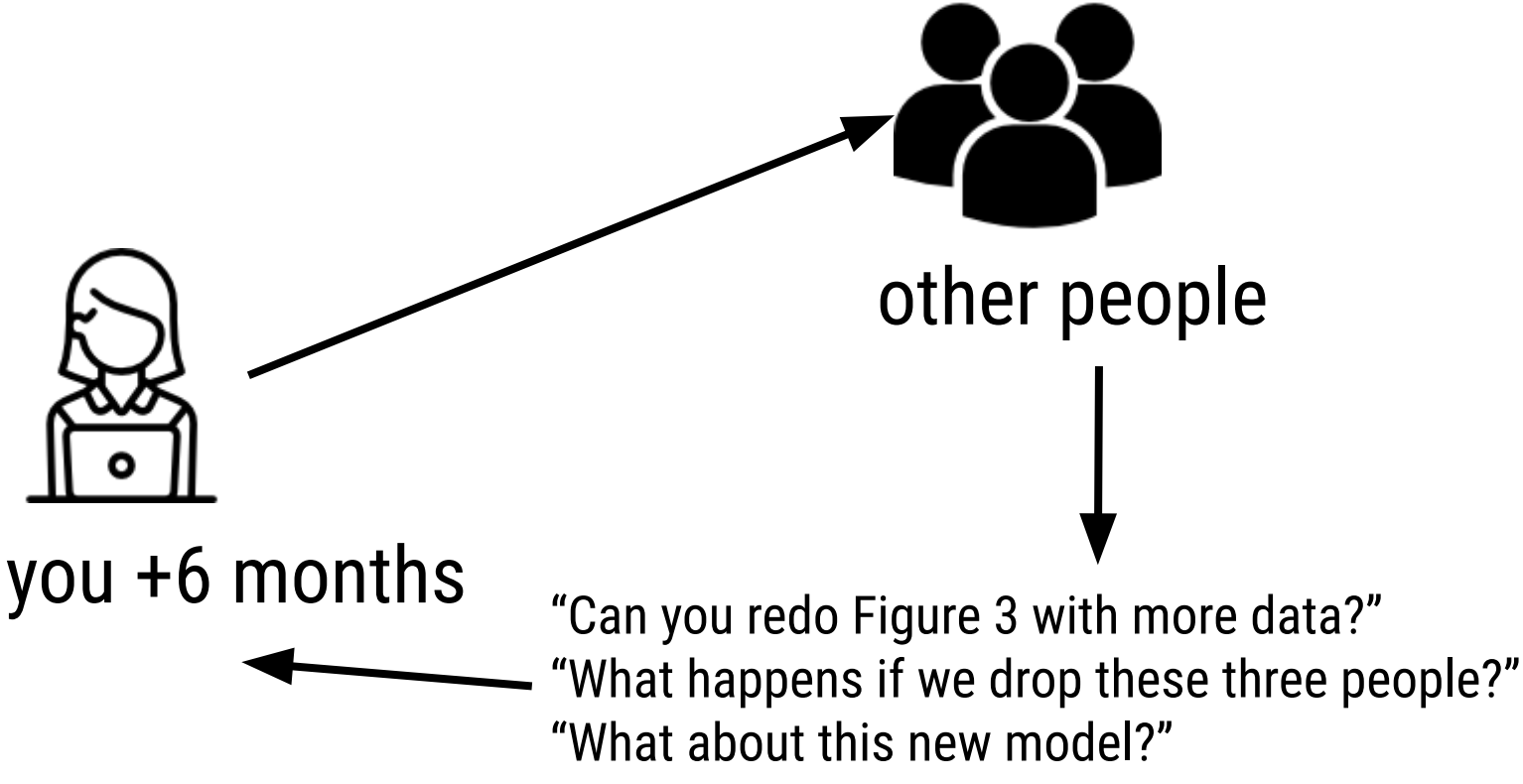
Data Science Files Should Be...

- Easy to find
- Easy to share
- Easy to understand
- Easy to update

Your audience

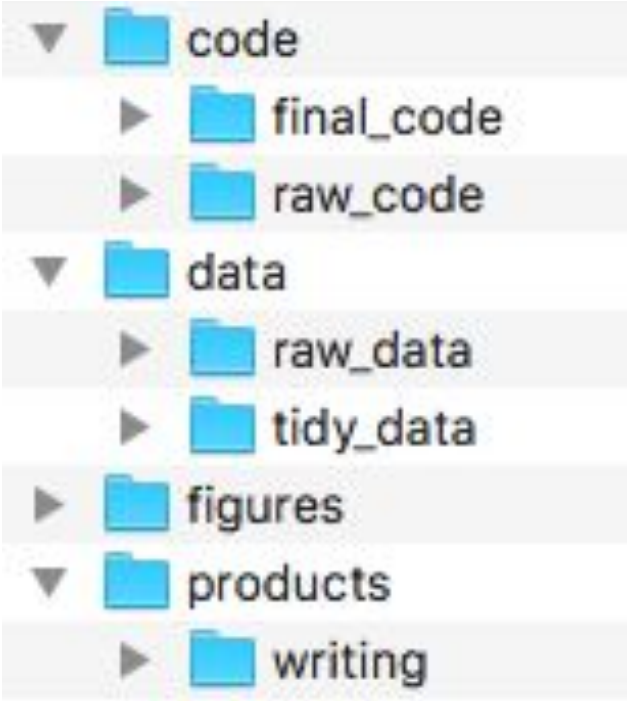


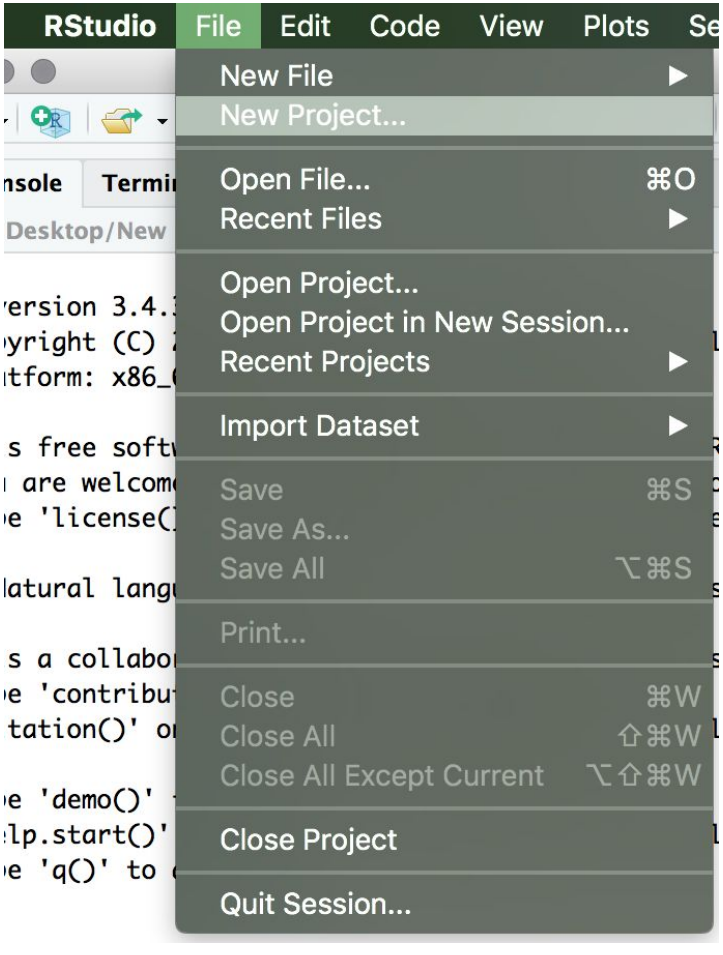
Why you are your own audience

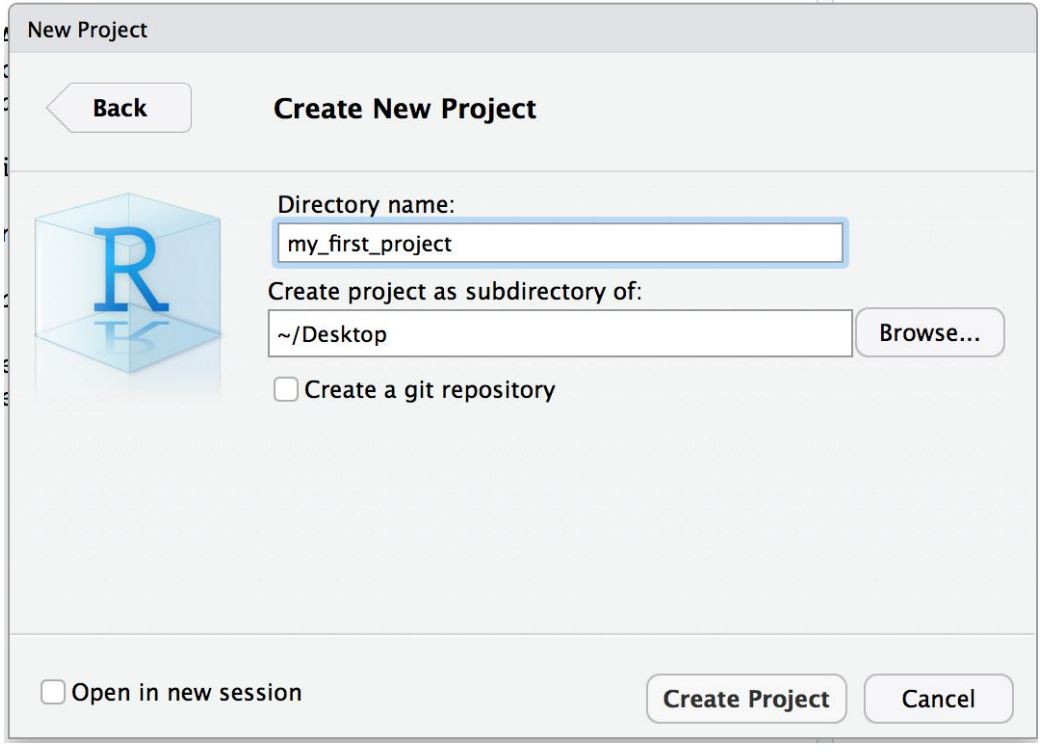


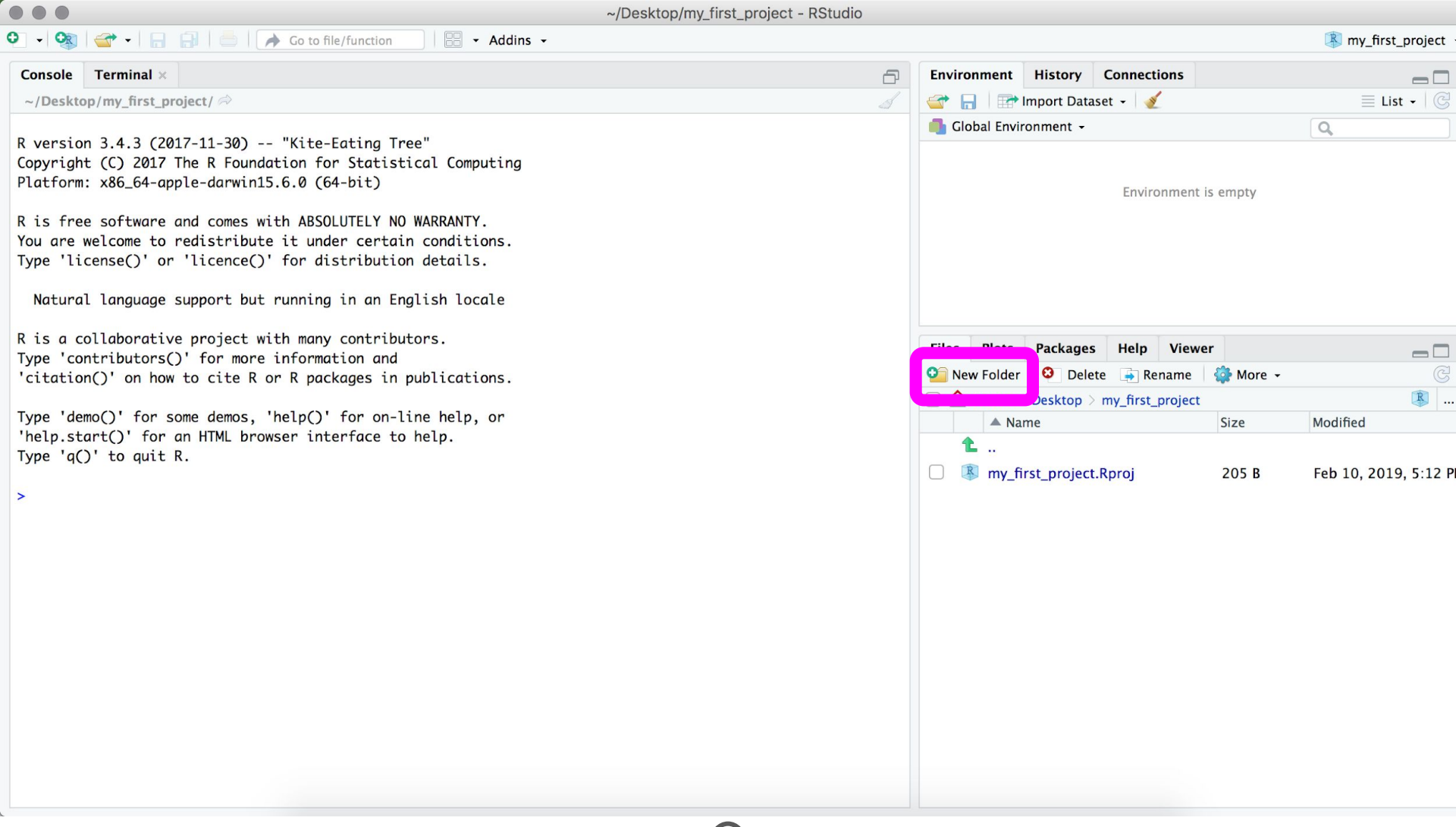
Why organize?

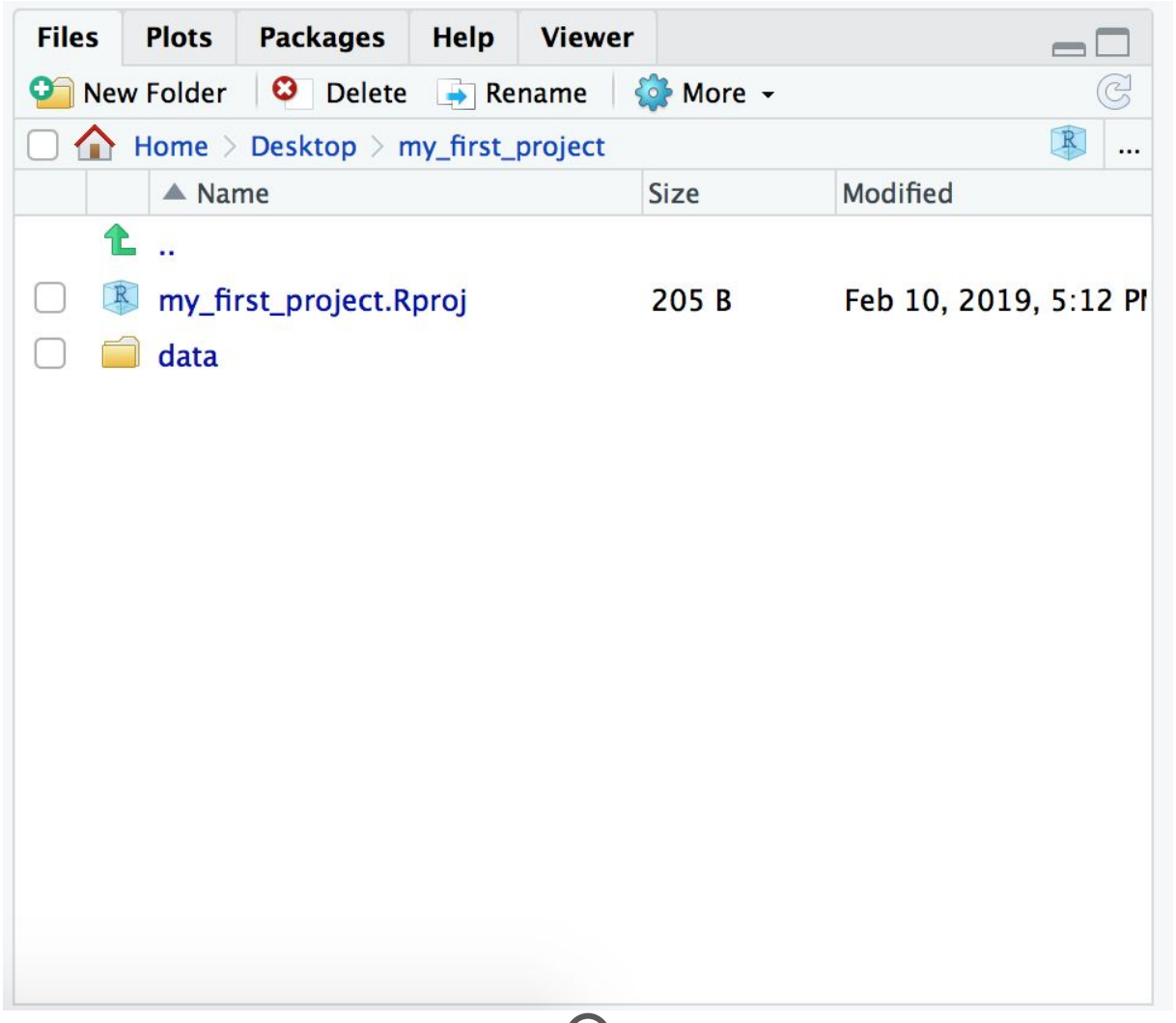
- It makes collaboration easier.
- It reduces the likelihood of making mistakes.
- It makes is a lot easier to go back to your analysis
- It shows transparency

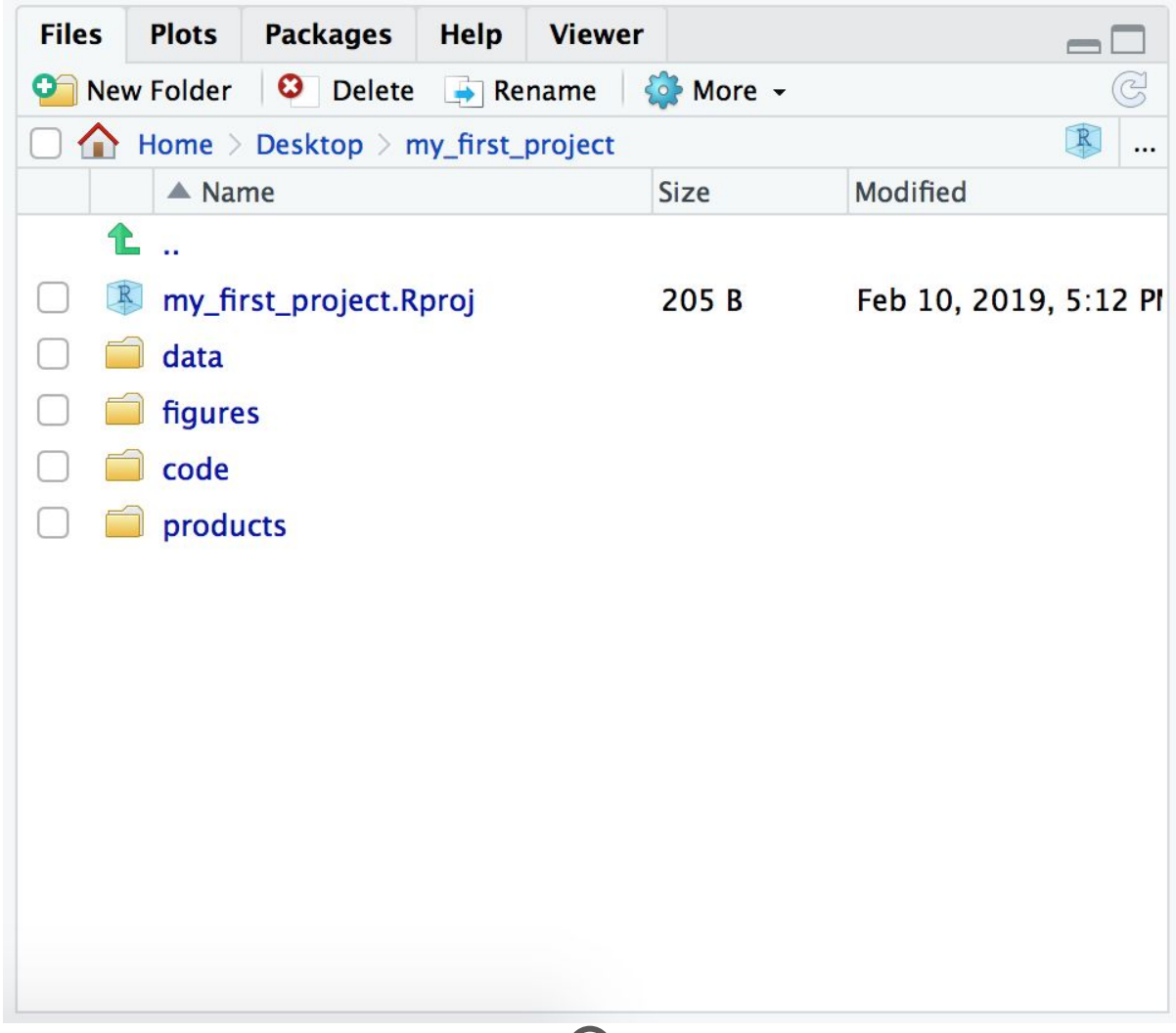

code
final_code
raw_code
data
raw_data
tidy_data
figures
products
writing

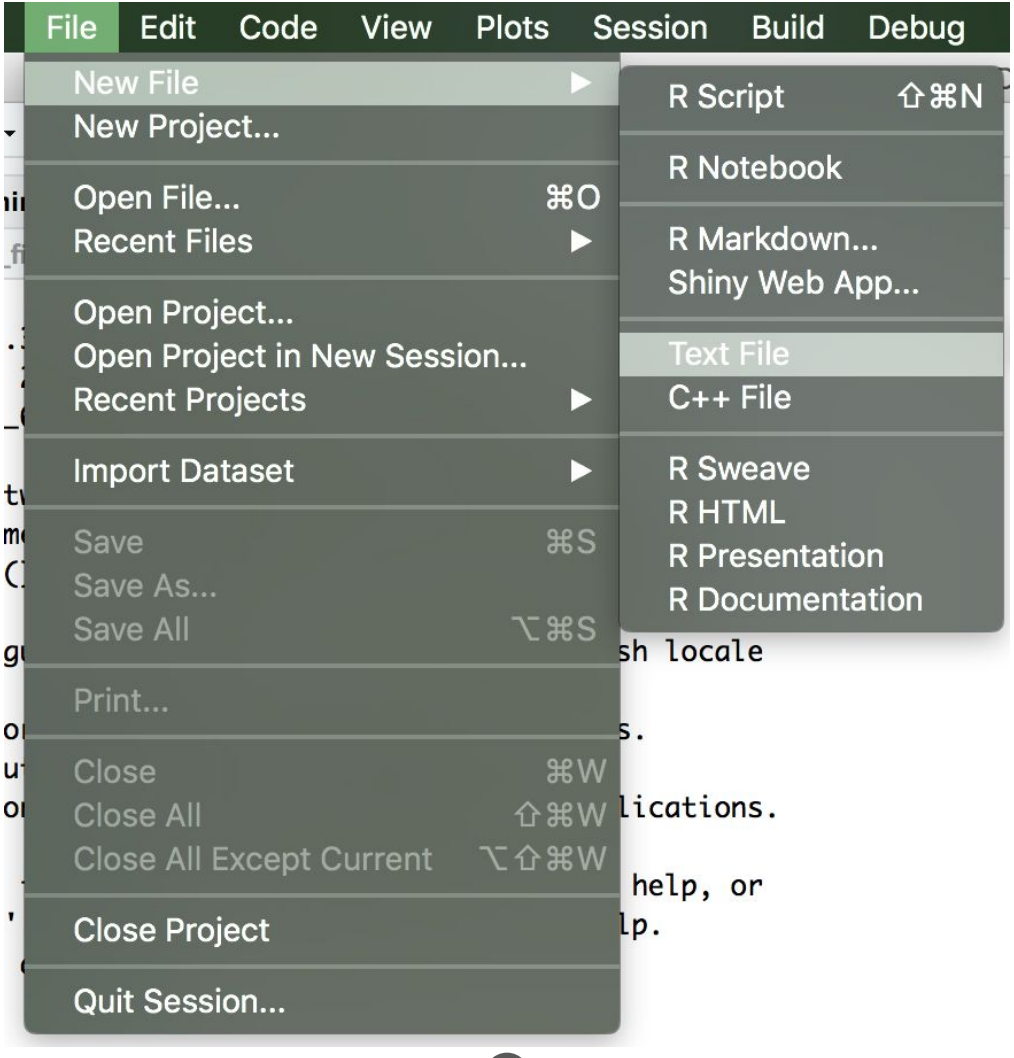

RStudio
File
Edit
Code
View
Plots
New File
New Project...
Open File... ⌘O
Recent Files
Open Project...
Open Project in New Session...
Recent Projects
Import Dataset
Save ⌘S
Save As...
Save All ⌥⌘S
Print...
Close ⌘W
Close All ⇧⌘W
Close All Except Current ⌥⇧⌘W
Close Project
Quit Session...

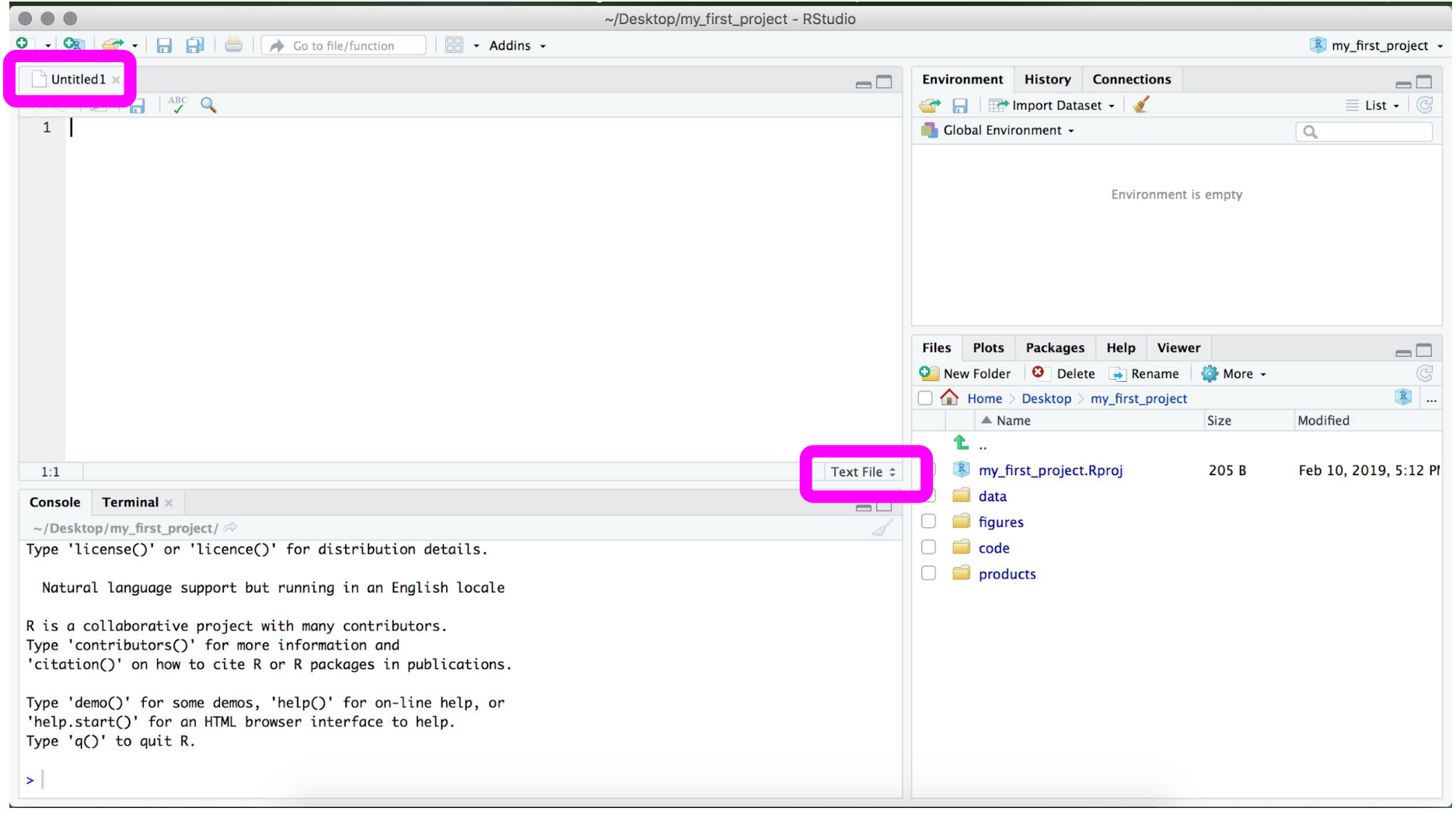

New Project
Back
Create New Project
Directory name:
my_first_project
Create project as subdirectory of:
~/Desktop
Browse...
Create a git repository
Open in new session
Create Project
Cancel

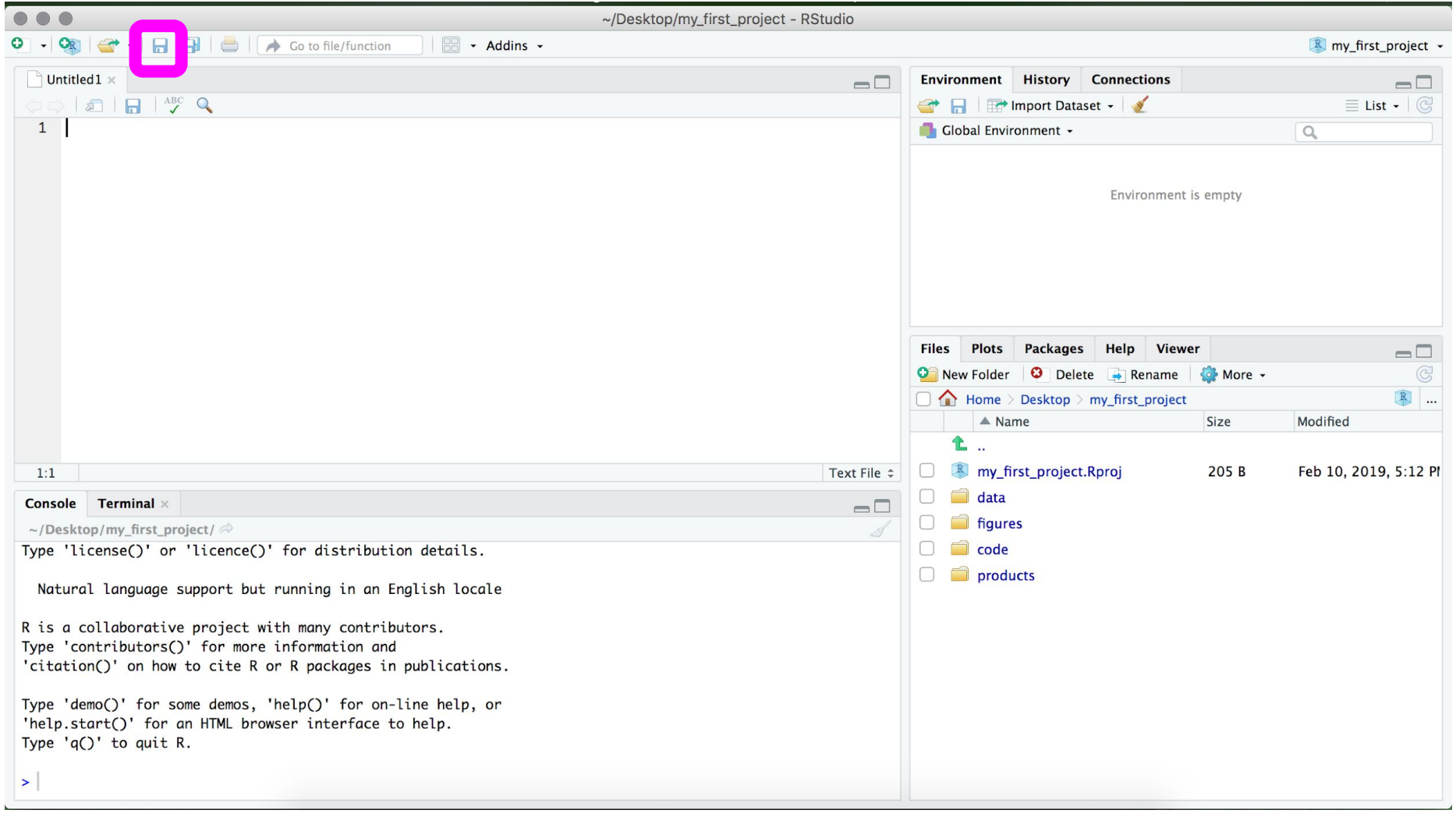

~/Desktop/my_first_project - RStudio
Go to file/function
Addins
my_first_project
Console
Terminal
~/Desktop/my_first_project/
R version 3.4.3 (2017-11-30) -- "Kite-Eating Tree"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Packages
Help
Viewer
New Folder
Delete
Rename
More
Desktop
my_first_project
Name
Size
Modified
..
my_first_project.Rproj
205 B
Feb 10, 2019, 5:12 PM

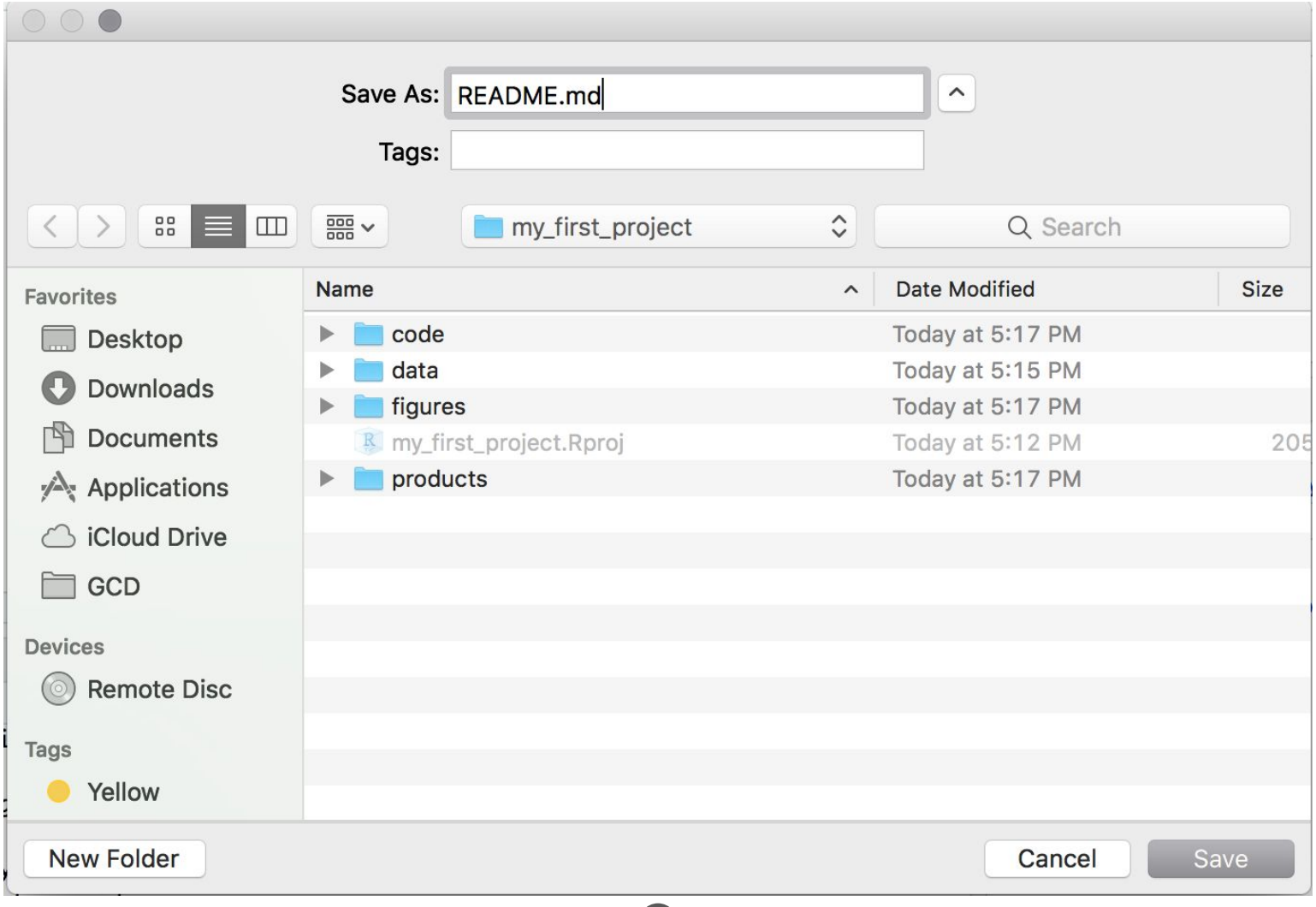

Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Desktop
my_first_project
Name
Size
Modified
..
my_first_project.Rproj
205 B
Feb 10, 2019, 5:12 PM
data

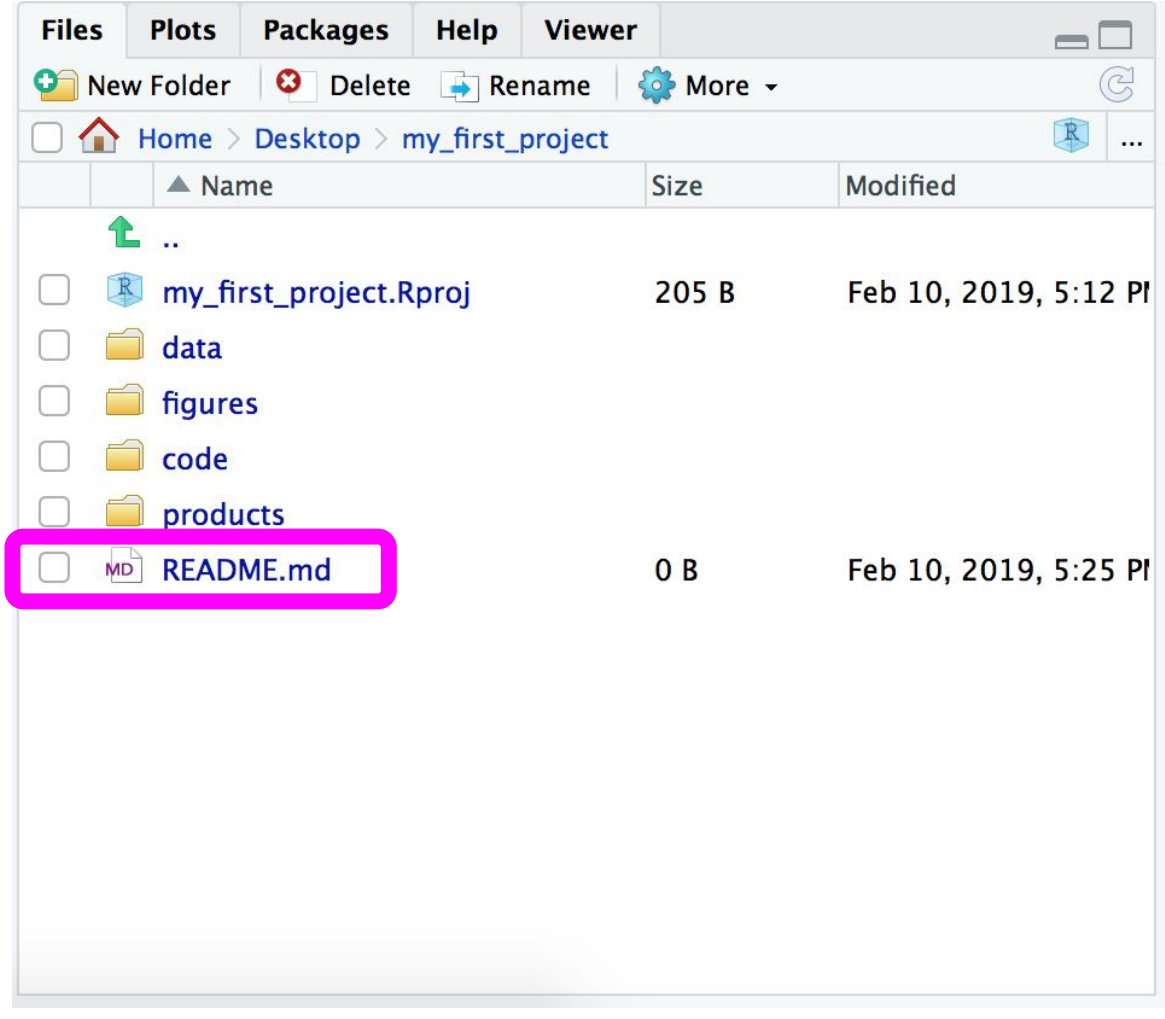

Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Desktop
my_first_project
Name
Size
Modified
..
my_first_project.Rproj
205 B
Feb 10, 2019, 5:12 PM
data
figures
code
products


File
Edit
Code
View
Plots
Session
Build
Debug
New File
New Project...
Open File... ⌘O
Recent Files
Open Project...
Open Project in New Session...
Recent Projects
Import Dataset
Save ⌘S
Save As...
Save All ⌥⌘S
Print...
Close ⌘W
Close All ⇧⌘W
Close All Except Current ⌥⇧⌘W
Close Project
Quit Session...
R Script ⇧⌘N
R Notebook
R Markdown...
Shiny Web App...
Text File
C++ File
R Sweave
R HTML
R Presentation
R Documentation


~/Desktop/my_first_project - RStudio
Go to file/function
Addins
my_first_project
Untitled1
1
1:1
Text File
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Desktop
my_first_project
Name
Size
Modified
..
my_first_project.Rproj
205 B
Feb 10, 2019, 5:12 PM
data
figures
code
products
Console
Terminal
~/Desktop/my_first_project/
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>


~/Desktop/my_first_project - RStudio
Go to file/function
Addins
my_first_project
Untitled1
1
1:1
Text File
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Desktop
my_first_project
Name
Size
Modified
..
my_first_project.Rproj
205 B
Feb 10, 2019, 5:12 PM
data
figures
code
products
Console
Terminal
~/Desktop/my_first_project/
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>


Save As: README.md
Tags:
my_first_project
Search
Favorites
Desktop
Downloads
Documents
Applications
iCloud Drive
GCD
Devices
Remote Disc
Tags
Yellow
Name
Date Modified
Size
code
Today at 5:17 PM
data
Today at 5:15 PM
figures
Today at 5:17 PM
my_first_project.Rproj
Today at 5:12 PM
205
products
Today at 5:17 PM
New Folder
Cancel
Save


Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Desktop
my_first_project
Name
Size
Modified
..
my_first_project.Rproj
205 B
Feb 10, 2019, 5:12 PM
data
figures
code
products
README.md
0 B
Feb 10, 2019, 5:25 PM

README.md

This is the README file for my_first_project

Last updated: 02-Mar-2018

The folders in this project are:

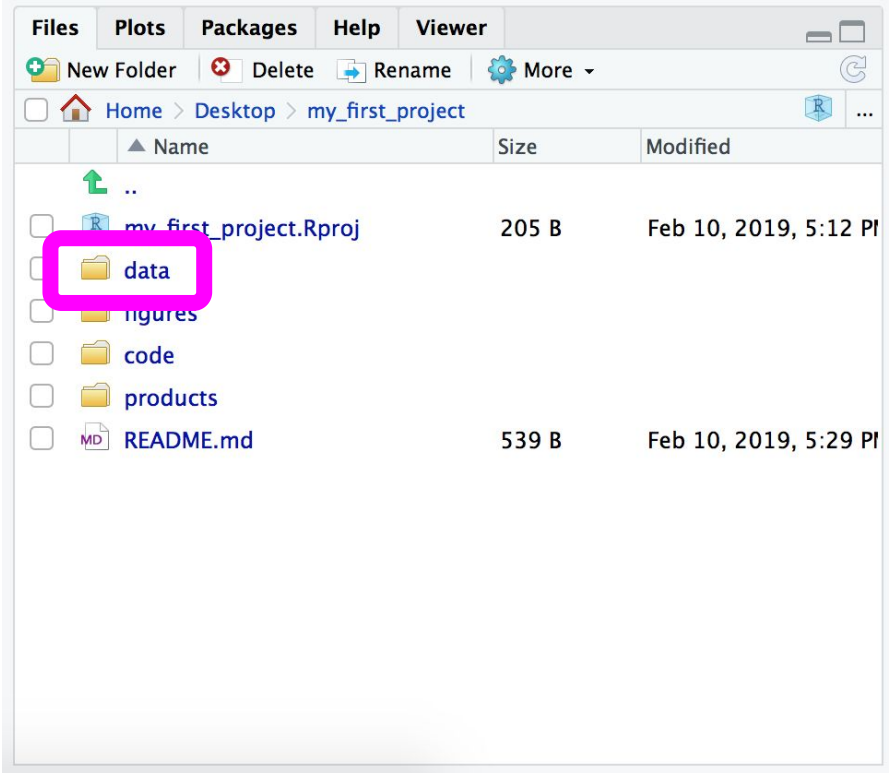
- *data* - is the folder where you can find all the collected data.
- *figures* - is where you can find all the plots, data pictures, and other images.
- *code* - is where you can find code files for collecting, cleaning up, or analyzing data.
- *products* - is where you can find reports, presentations, or products

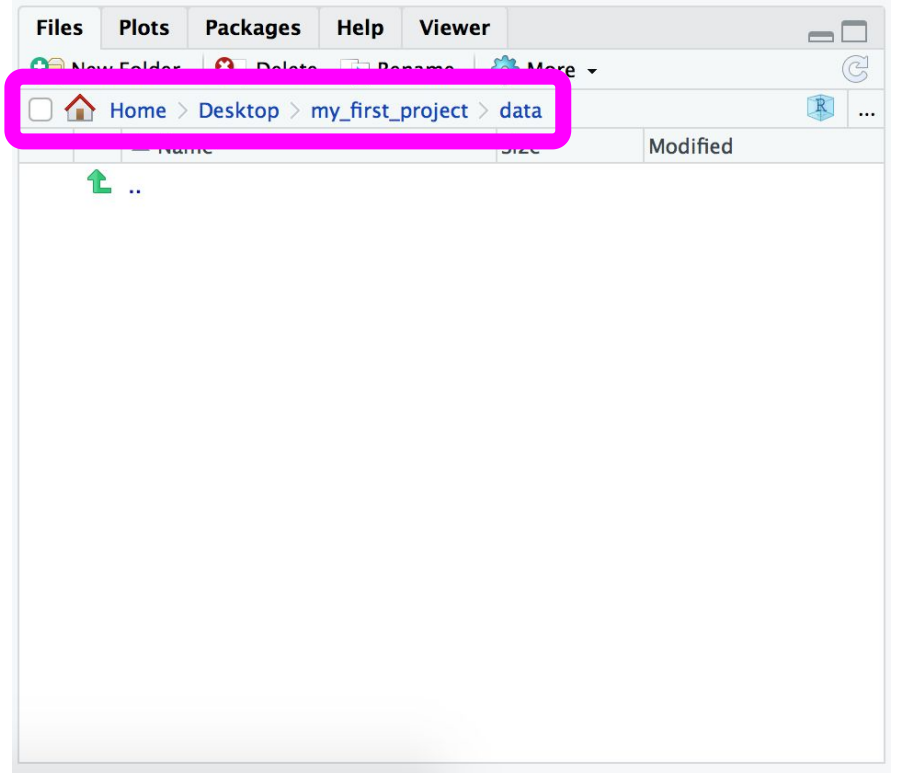
Data on crime is obtained from International Crime Data collected between 2015-2018 and is publicly available. Data on happiness is collected from the Survey of International Happiness.

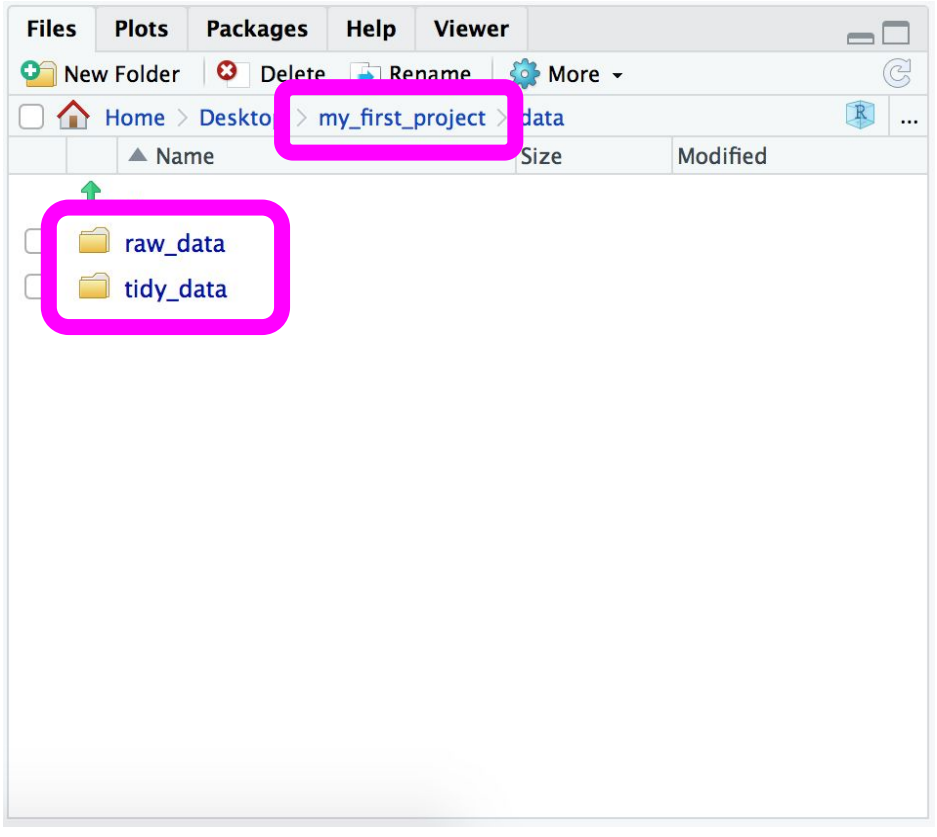
Contributors:

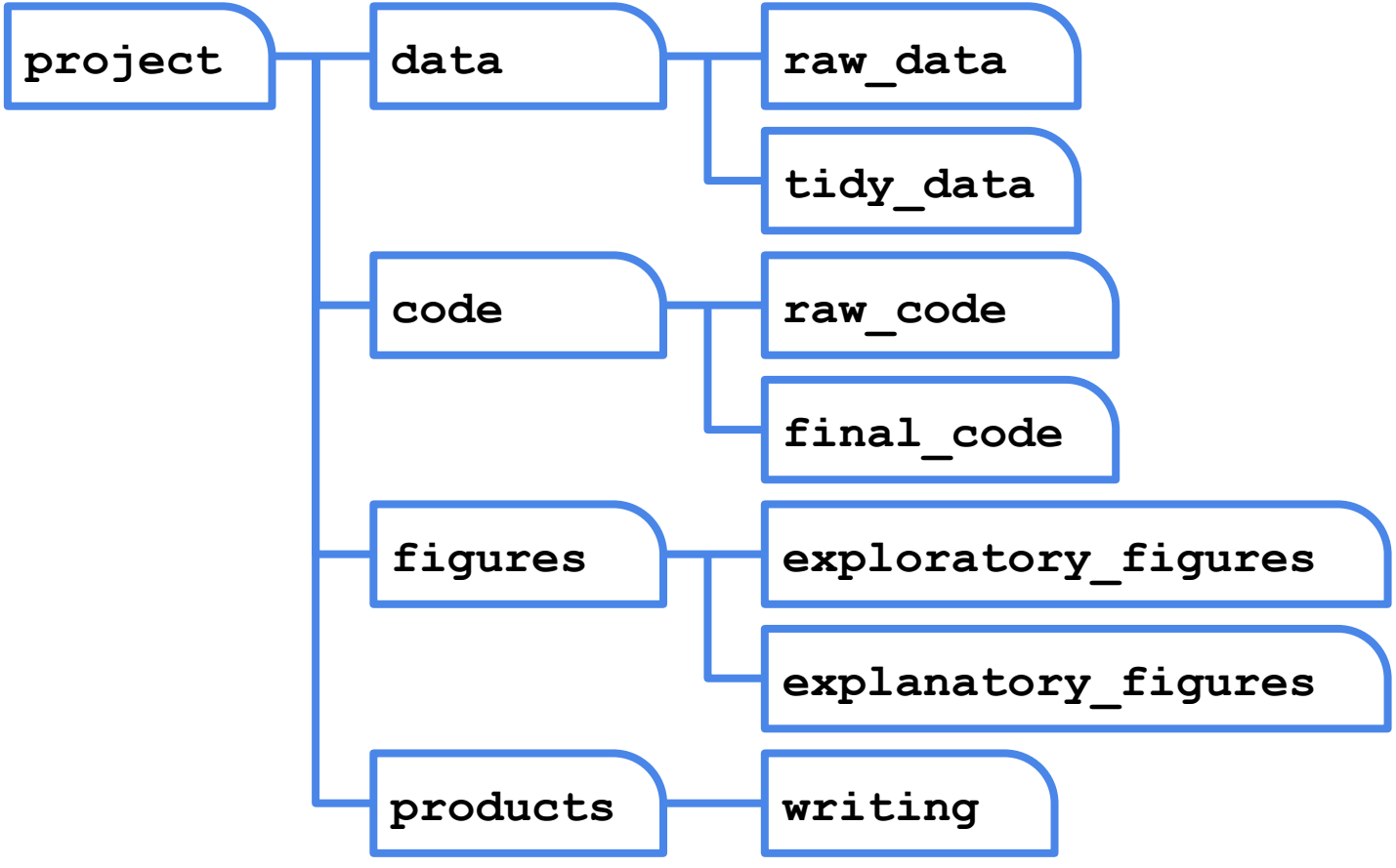
- Jane Everyday Doe, jane.everyday.doe@gmail.com
- John Everyday Doe, john.everyday.doe@gmail.com

Cite: Doe, J, and Doe, J, Sample Analysis Using Sample Data, Working Paper, 2018


Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Desktop
my_first_project
Name
Size
Modified
..
my_first_project.Rproj
205 B
Feb 10, 2019, 5:12 PM
data
figures
code
products
README.md
539 B
Feb 10, 2019, 5:29 PM


Files
Plots
Packages
Help
Viewer
Home
Desktop
my_first_project
data
Modified
..


Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Deskto
my_first_project
data
Name
Size
Modified
raw_data
tidy_data


project
data
raw_data
tidy_data
code
raw_code
final_code
figures
exploratory_figures
explanatory_figures
products
writing

Key principles of file naming for data science projects:

- Machine readable
- Human readable
- Be nicely ordered

Source: Jenny Bryan

Bad Naming	Good Naming
2013 my report.md	2013_my_report.md
malik's_report.md	maliks_report.md
01_zoë_report.md	01_zoe_report.md
AdamHooverReport.md	adam-hoover-report.md
executivereportpepsiv1.md	executive_report_pepsi_v1.md

2018_jan_sales_cust001_prod001.md
2017_mar_sales_cust001_prod001.md
2016_may_sales_cust001_prod008.md
2017_jan_sales_cust120_prod007.md
2015_oct_sales_cust034_prod001.md
2015_oct_sales_cust034_prod002.md

Year	Month	Type	Customer ID	Product ID
2018	jan	sales	001	001
2017	mar	sales	001	001
2016	may	sales	001	008
2017	jan	sales	120	007
2015	oct	sales	034	001
2015	oct	sales	034	002

Which one is better?

analysis.R
or
2019-exploratory_analysis_crime.R?

Which one is better?

05-21-2017-analysis-cust001.R
or
2017-05-21-analysis-cust001.R?

Report01_cust12_prod03.md
Report02_cust12_prod03.md
Report03_cust12_prod03.md
Report04_cust12_prod03.md

Bad Naming	Good Naming
1_consumer_analysis.md	001_consumer_analysis.md
reg_1_company_data.R	reg_01_company_data.R

Summarizing: Organizing your data analysis

Getting and Cleaning Data