Research workflow: Stata, Github, Overleaf and R

Workshop on Automation of the Research Process

Alex Hollingsworth, Ohio State University and NBER

Be replicable

Have a workflow

Invest in reducing the cost of errors, future, work.

Invest in future you.

Specific tools do not matter. Buy in matters

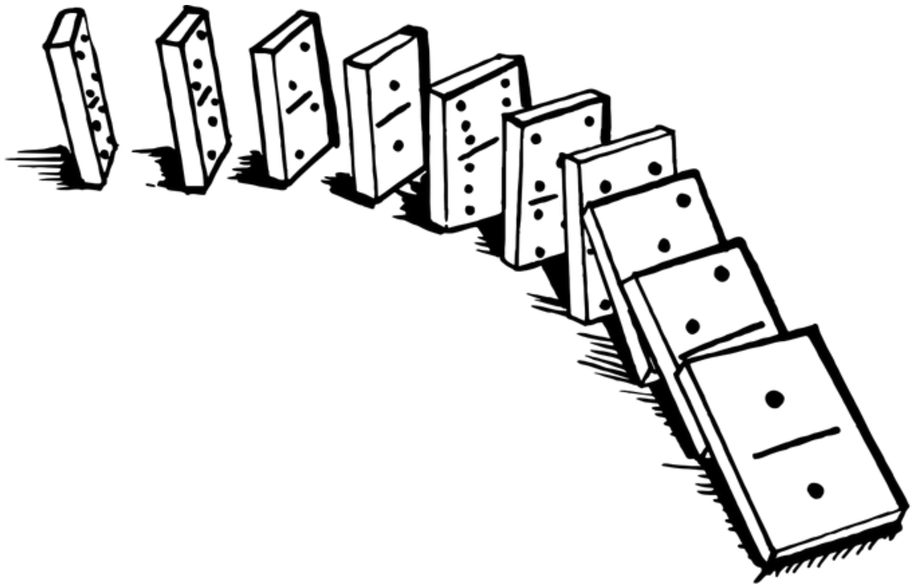
- “Imagine if you were reading a great theory paper that had elegant theorems, but then you went to go look at the proofs everything was messy and incomplete…It would make you rethink the validity of the original paper… The code underlying [an empirical] paper, is a lot like the proof [underlying a theory paper].”

- “We want to get the right answer. And having something that’s organized makes [this] more likely”

Julian Reif, Associate Professor of Finance and Economics at Gies College of Business, University of Illinois

Keystone habit

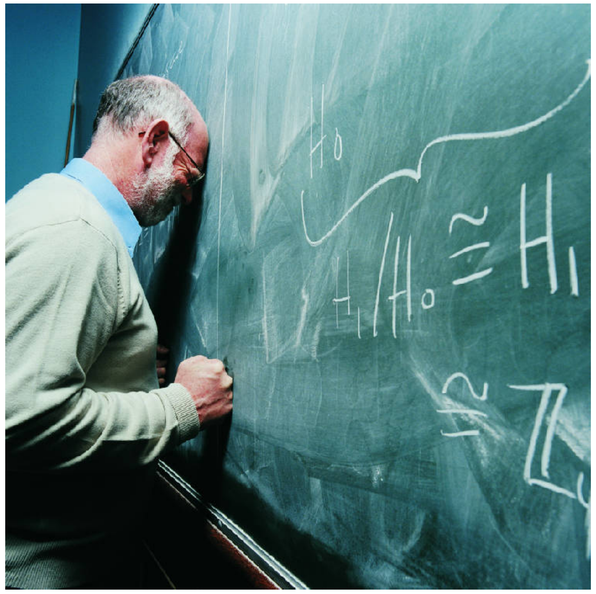
- **Work in a reproducible manner that can be easily shared with others**
  - Reduces the probability of error
  - Improves the “scientific” process
  - Practice ethical research
    - Reduce temptation to specification hunt
    - p-hack
    - file-drawer results
  - Improves collaboration
    - With your current co-authors
    - Future researchers
    - Smooths team transitions
  - Lends credibility to your results

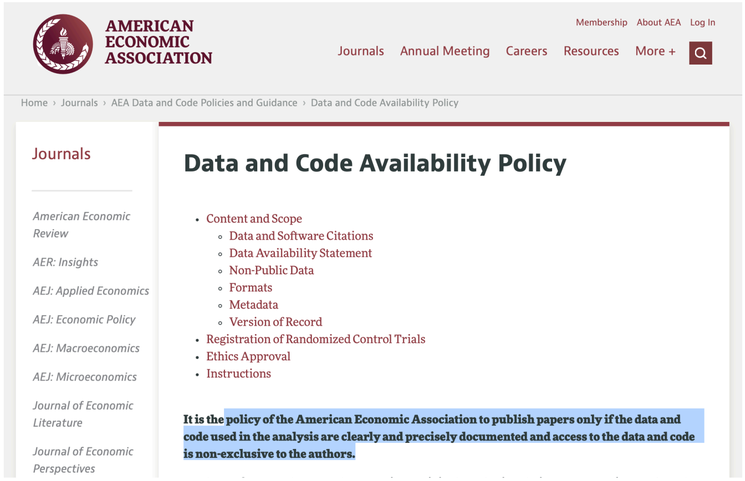


- A thoughtful, partially automated, workflow is essential to achieving this goal

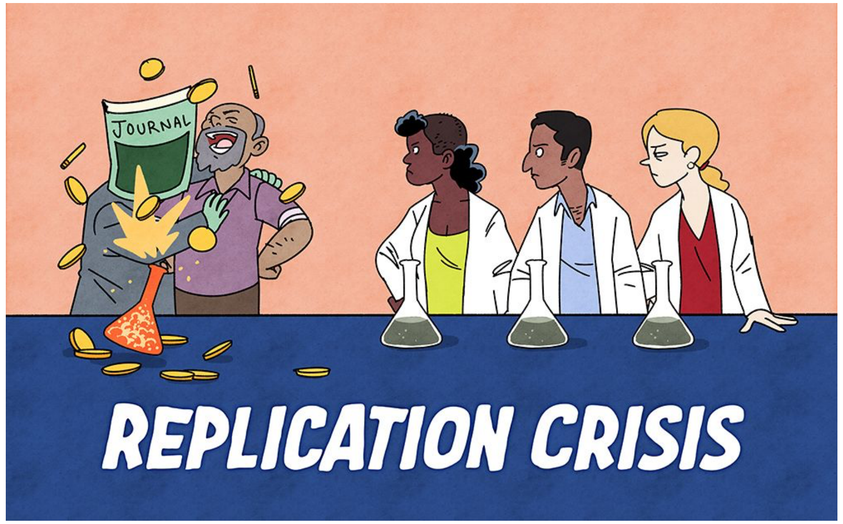
Why should you care about workflow?

Future you





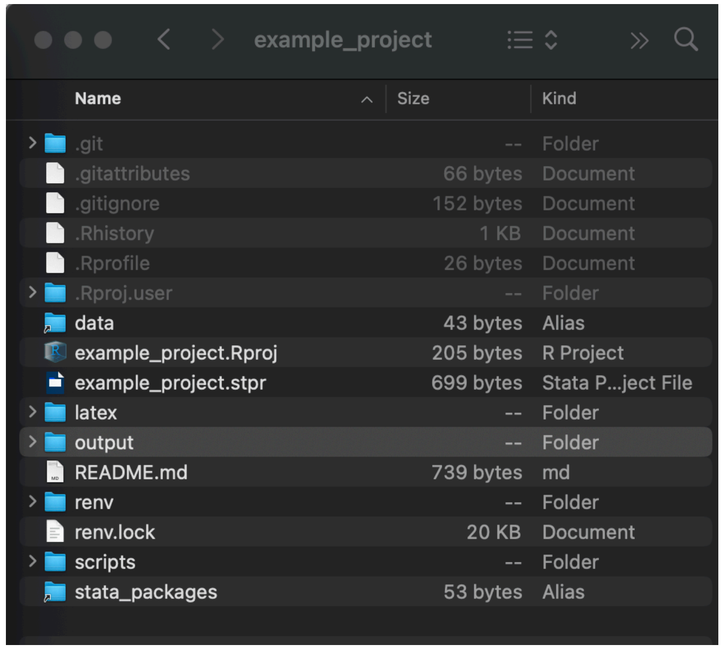
Future research



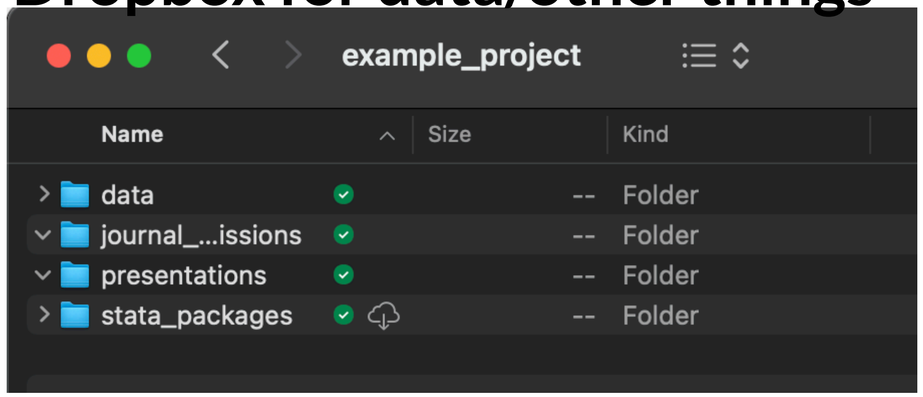
Set-up

Visual overview

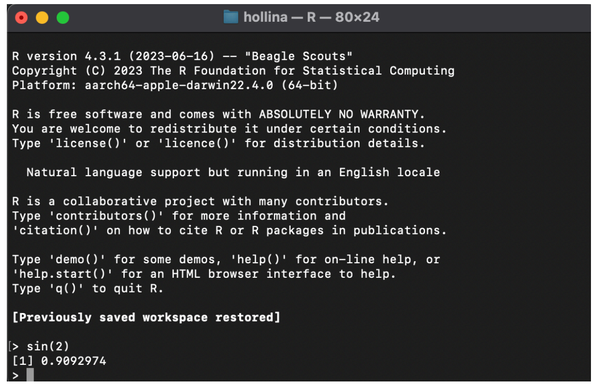
GitHub for code



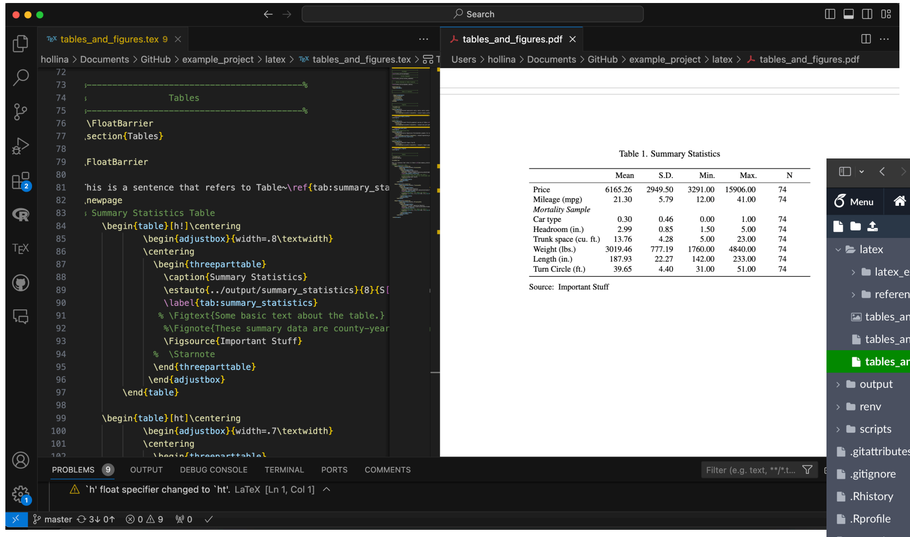
Dropbox for data/other things



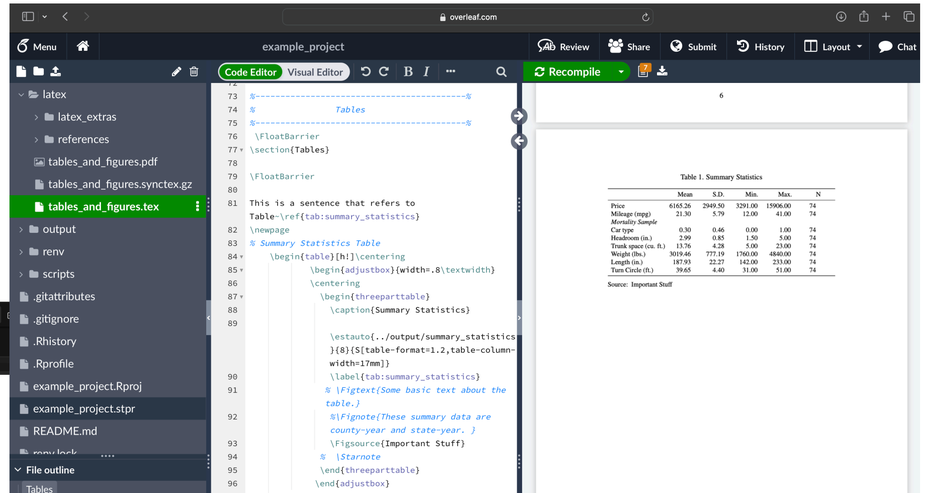
Coding/analysis on server



Local coding/writing using VSCode

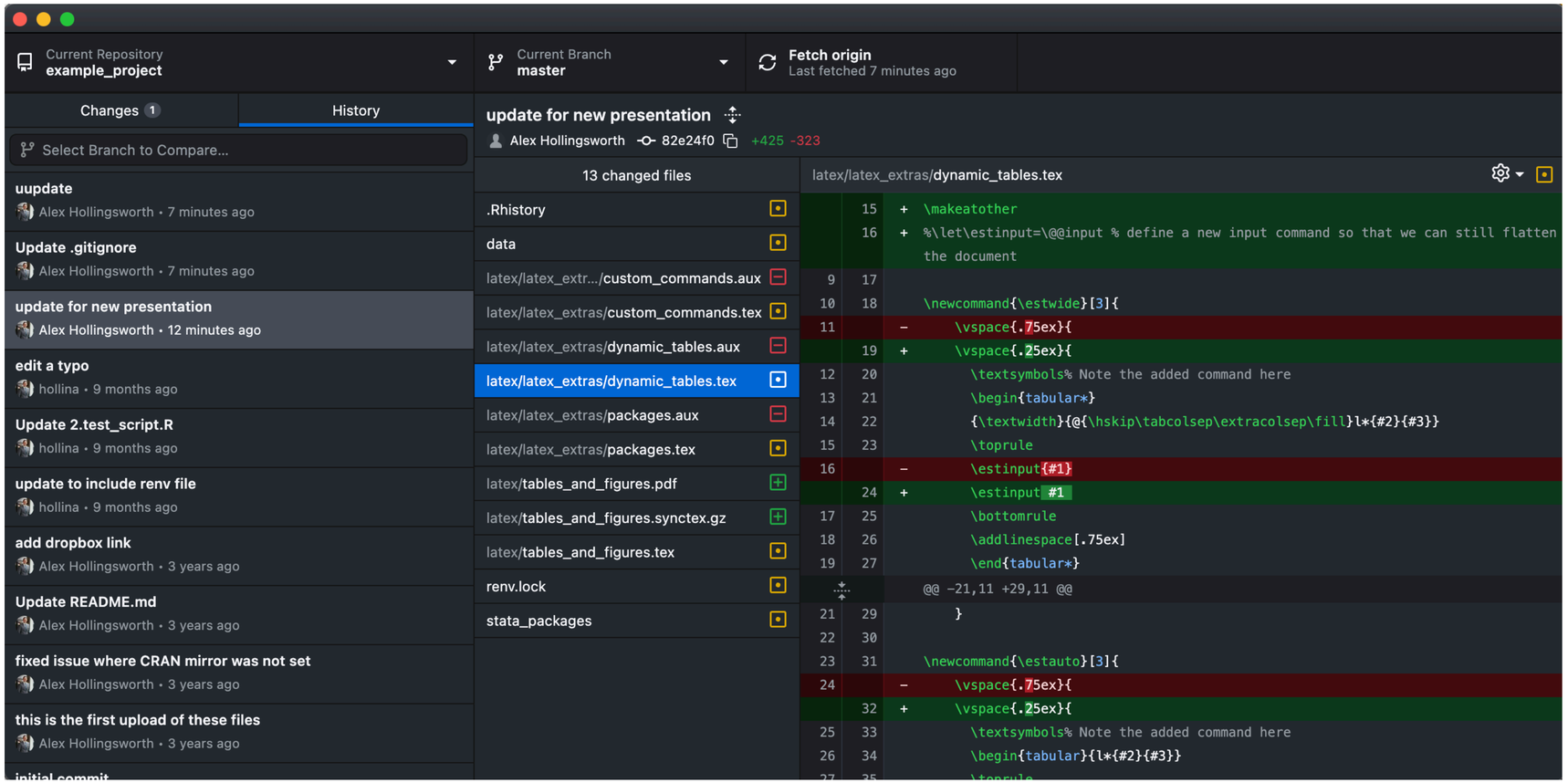


Online writing using overleaf



GitHub is the “glue”

- It’s more complicated than it needs to be for your purposes
- Think of it as an extra save button you need to click
  - fancy save
  - Handles merging really well



Basic idea of GitHub

- You write/edit some code or a document
  - You click save. This saves locally.

Basic idea of GitHub

- You write/edit some code or a document
  - You click save. This saves locally.
- After a while, say you finish a do-file, need to end the work day,
  - You write a little message in your GitHub saying what you did
    - AKA commit

Basic idea of GitHub

- You write/edit some code or a document
  - You click save. This saves locally.
- After a while, say you finish a do-file, need to end the work day,
  - You write a little message in your GitHub saying what you did
- You check to see if anyone else has made changes
  - AKA fetch

Basic idea of GitHub

- You write/edit some code or a document
  - You click save. This saves locally.
- After a while, say you finish a do-file, need to end the work day,
  - You write a little message in your GitHub saying what you did
- You check to see if anyone else has made changes
- Then you “push” these changes to a server that can then be accessed on all your devices

Basic idea of GitHub

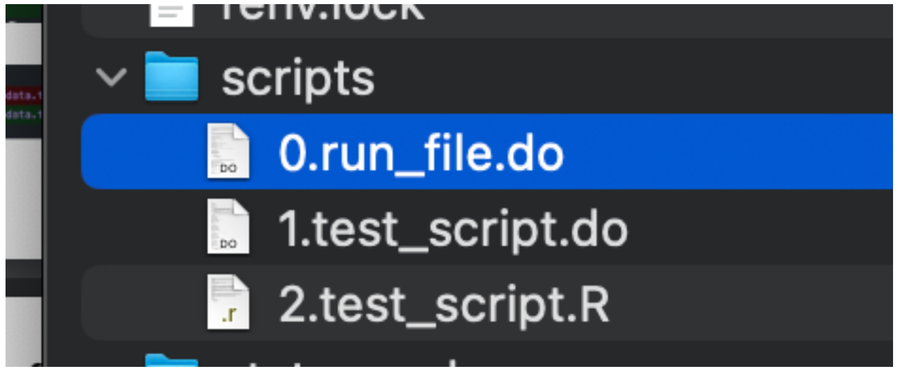
- You write/edit some code or a document
  - You click save. This saves locally.
- After a while, say you finish a do-file, need to end the work day,
  - You write a little message in your GitHub saying what you did
- You check to see if anyone else has made changes
- Then you “push” these changes to a server that can then be accessed on all your devices
- If there are any conflicts, you solve them via “merge”

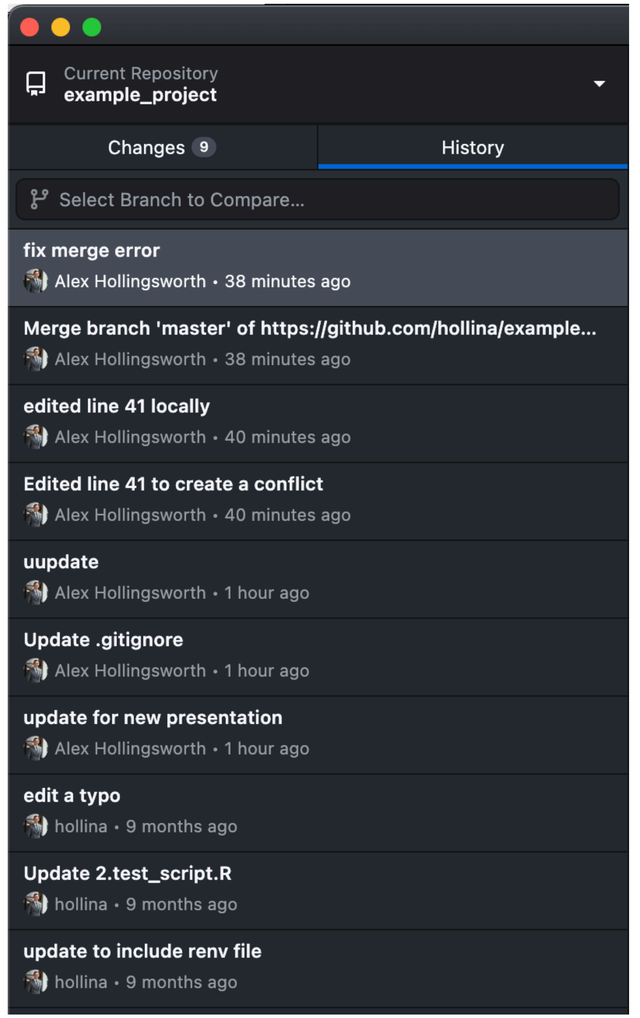
Basic idea of GitHub

- There is a detailed log kept of all changes made to code
- Everything that you want coarser version control for is stored using another system like dropbox

Main benefit: version control

- Only one file! No more of this `version_1245_final_for_real_this_time.do`





Main benefit: version control

latex/duke_endowment.tex

```
tud
2518  2516        \Fignote{* p $<$ 0.1, ** p $<$ 0.05, *** p $<$ 0.01.
2519  2517              Each coefficient comes from a separate Poisson pseudo maximum likelihood regression.
2520  2518              In columns 1 to 3, the unit of observation is a county-by-year of birth cell.
2521        -             In columns 4 to 6, the unit of observation is a birth county by birth year by follow-up
      year triplet.
      2519  +           In columns 4 to 6, the unit of observation is a birth county by birth year by follow-up
      age triplet.
```

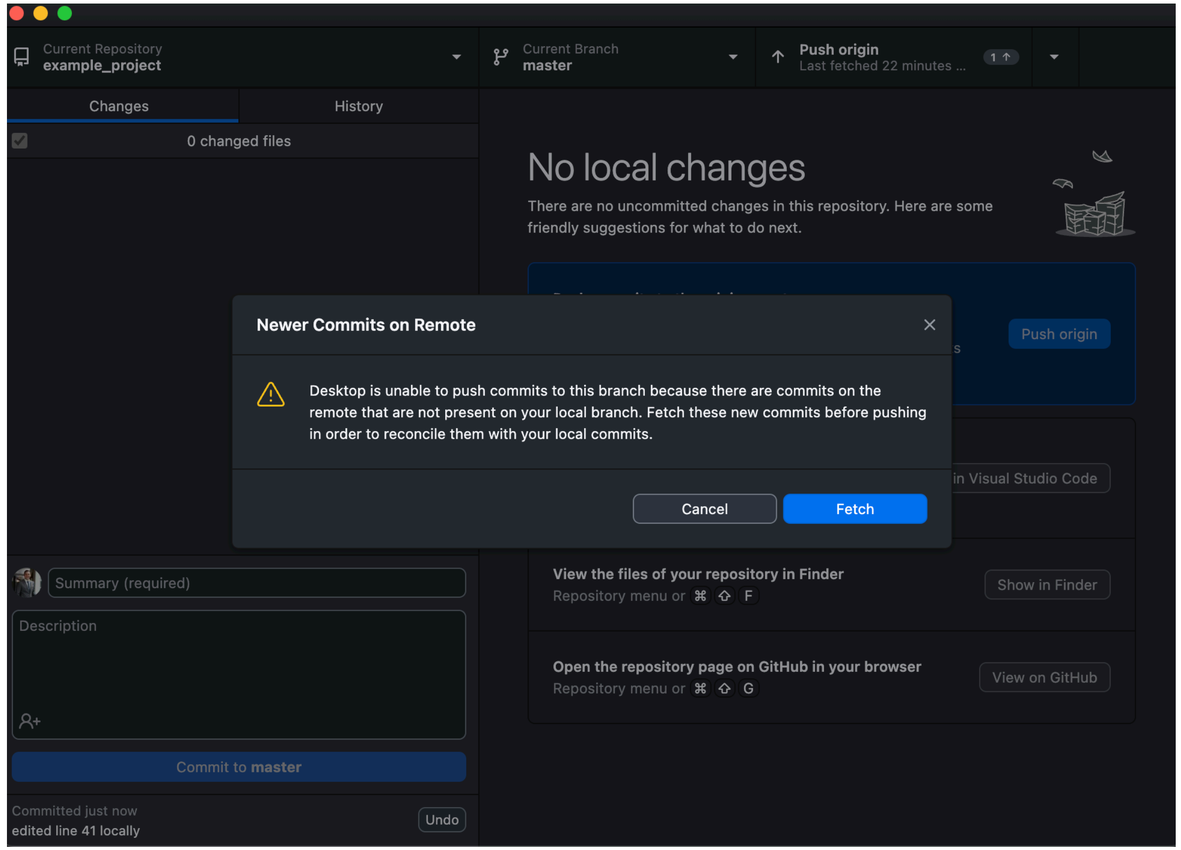
```
14  14
15      -  p_load(tidyverse, data.table)
        15 +  p_load(tidyverse, data.table, lubridate)
16  16
```

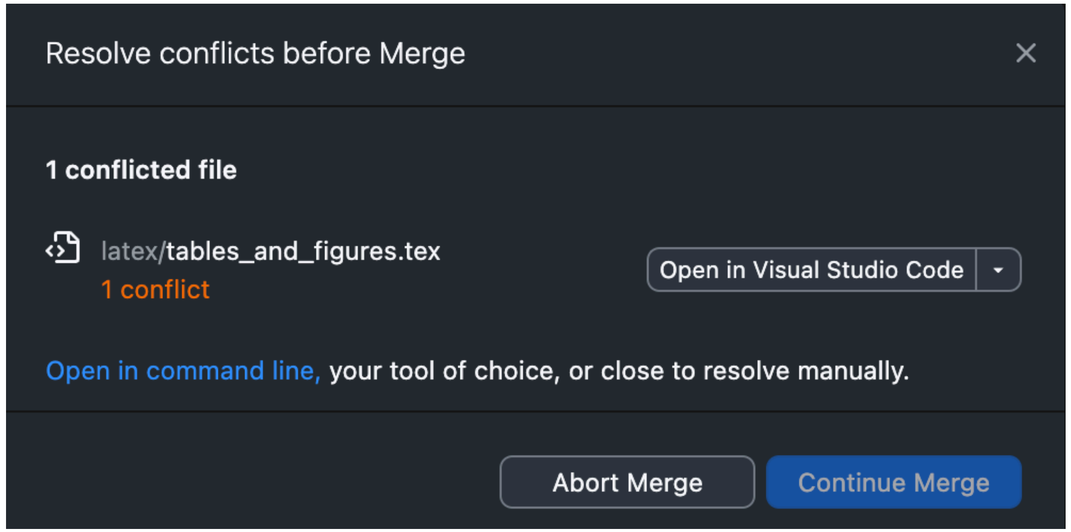
replication/analysis/output/appendix/table_hX_maternal_mortality_diff_specs.tex

```
                @@ -1,11 +1,11 @@
1   1           & \multicolumn{3}{c}{\shortstack{Number of Deaths}} & \multicolu
            \multicolumn{1}{c}{(1)} & \multicolumn{1}{c}{(2)} & \multicolumn
2   2       \addlinespace
3       -   \midrule \emph{A. Maternal mortality} &&&&&& \\ \addlinespace\hs
            &       -2.60          &      -11.87          &      -12.83
4       -   ~                     &      (8.39)          &      (7.33)
        3   + \midrule \emph{A. Maternal mortality} &&&&&& \\ \addlinespace\hs
            &       -3.60          &      -16.84*         &     -18.37**
        4   + ~                     &      (8.39)          &      (7.33)
5   5       \addlinespace\hspace{.5cm} Observations&\multicolumn{1}{c}{1,100
```

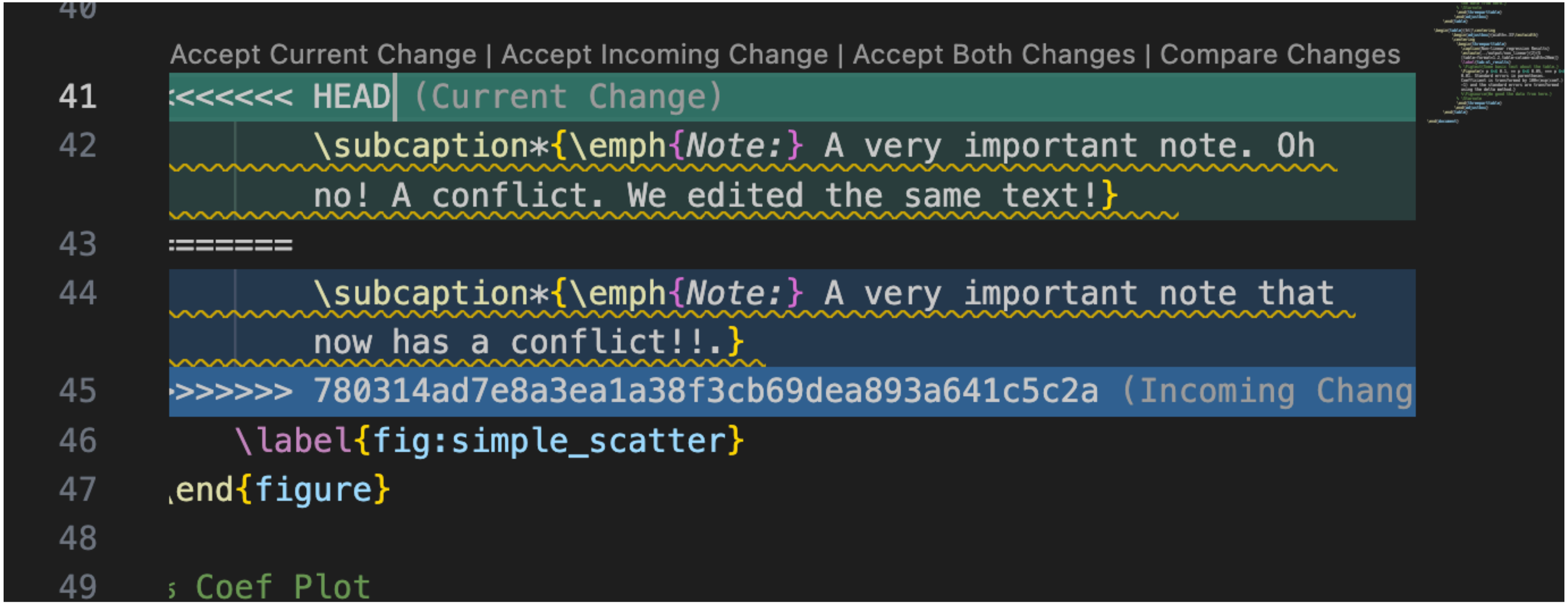
Second benefit: merge conflicts

- Say you and someone else edited the same document. Sound familiar?

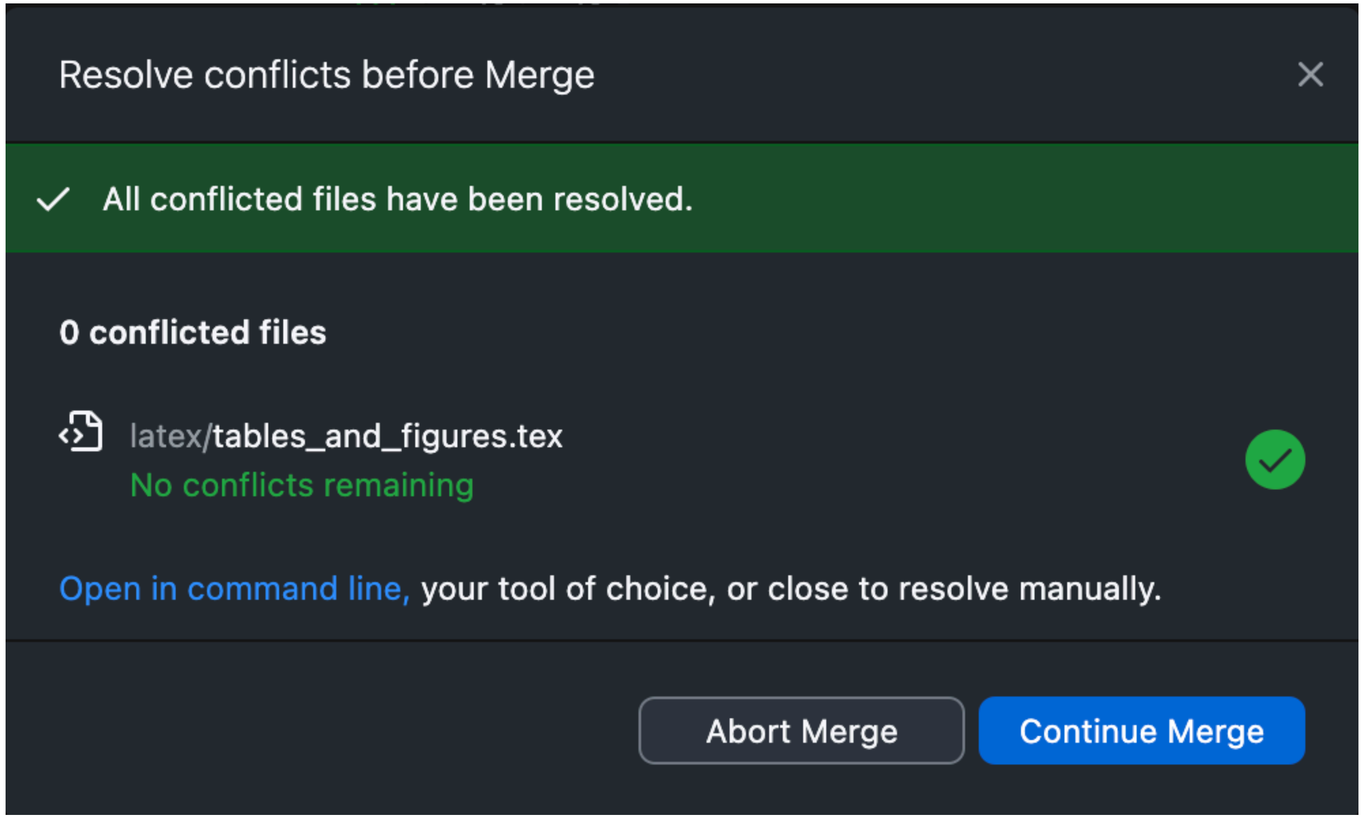




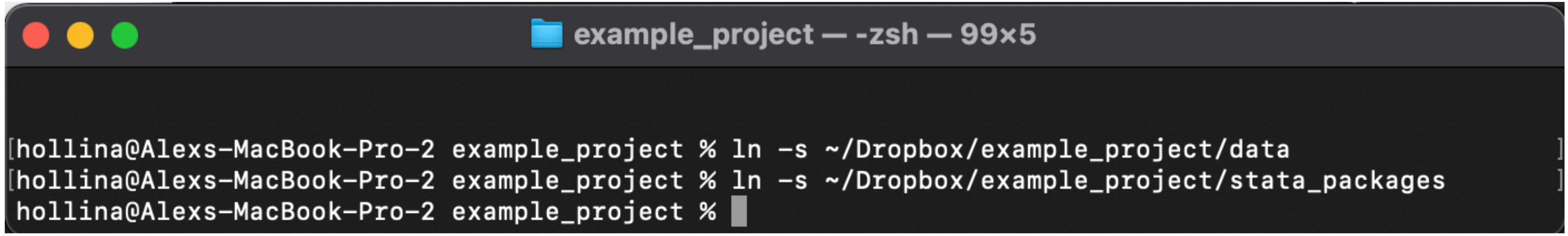
Second benefit: merge conflicts

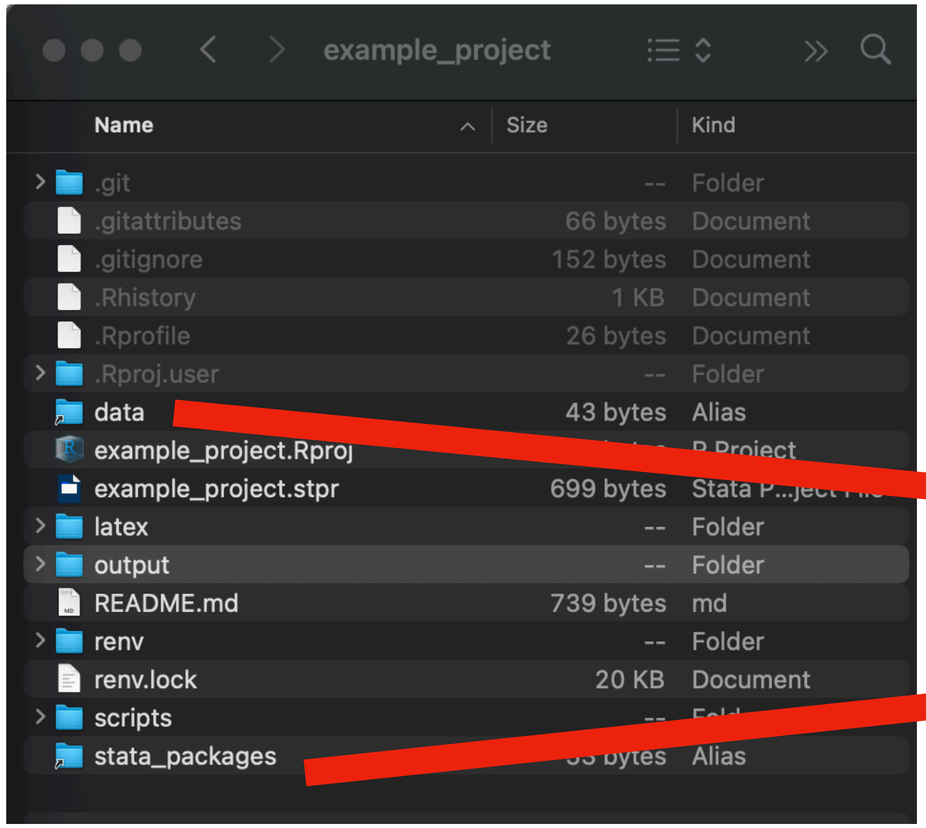


Second benefit: merge conflicts



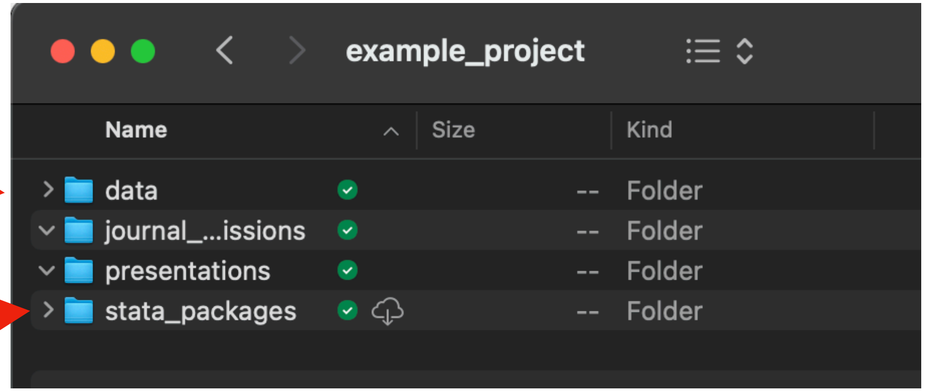
Symbolic links





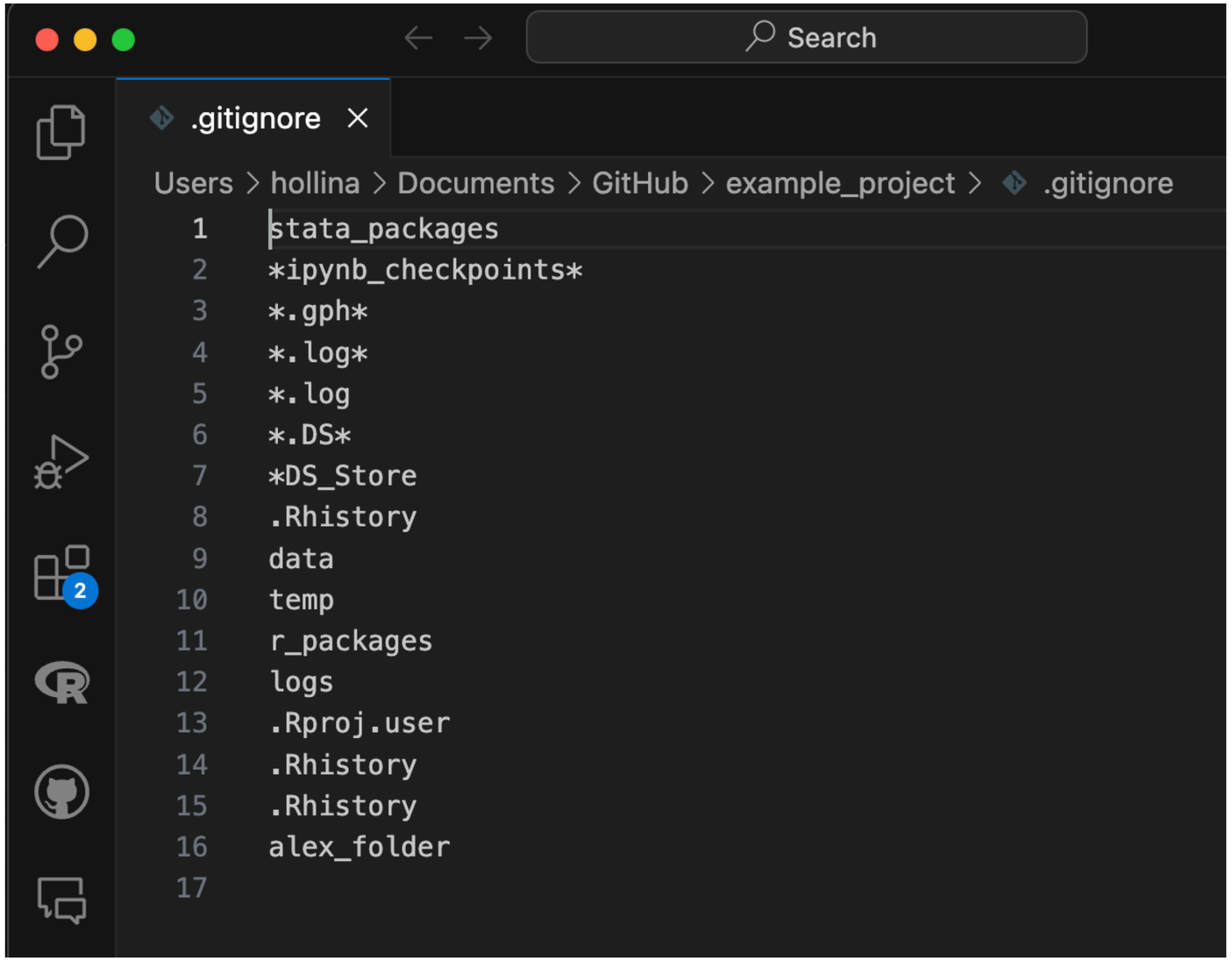
- This is basically a shortcut from your Dropbox to your GitHub

Dropbox for data/other things



GitIgnore

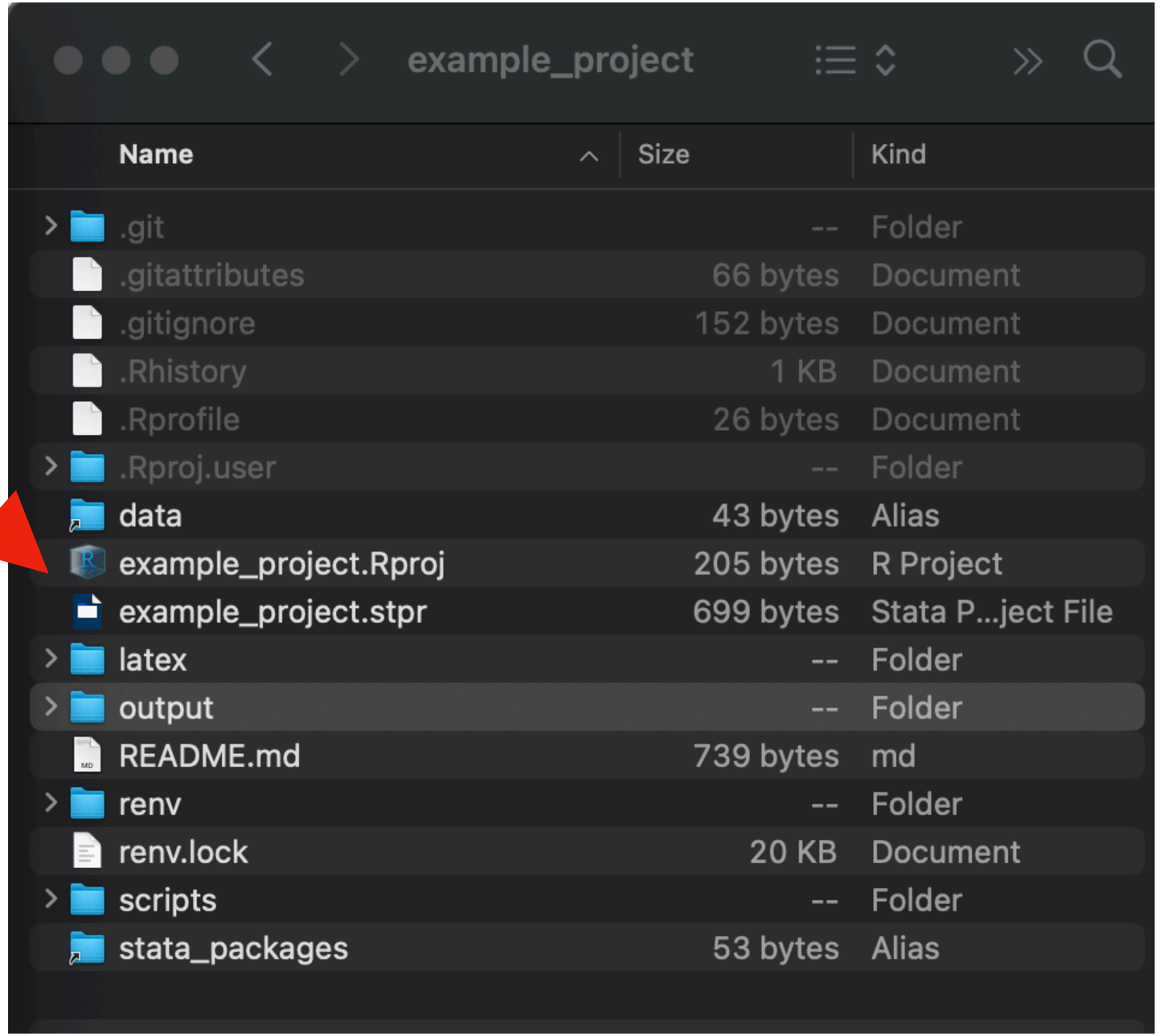
- List of files you don’t want GitHub to “track”
  - If you use symbolic links they can still be backed up via DropBox



Projects

- Think of this as an anchor to where your project is located

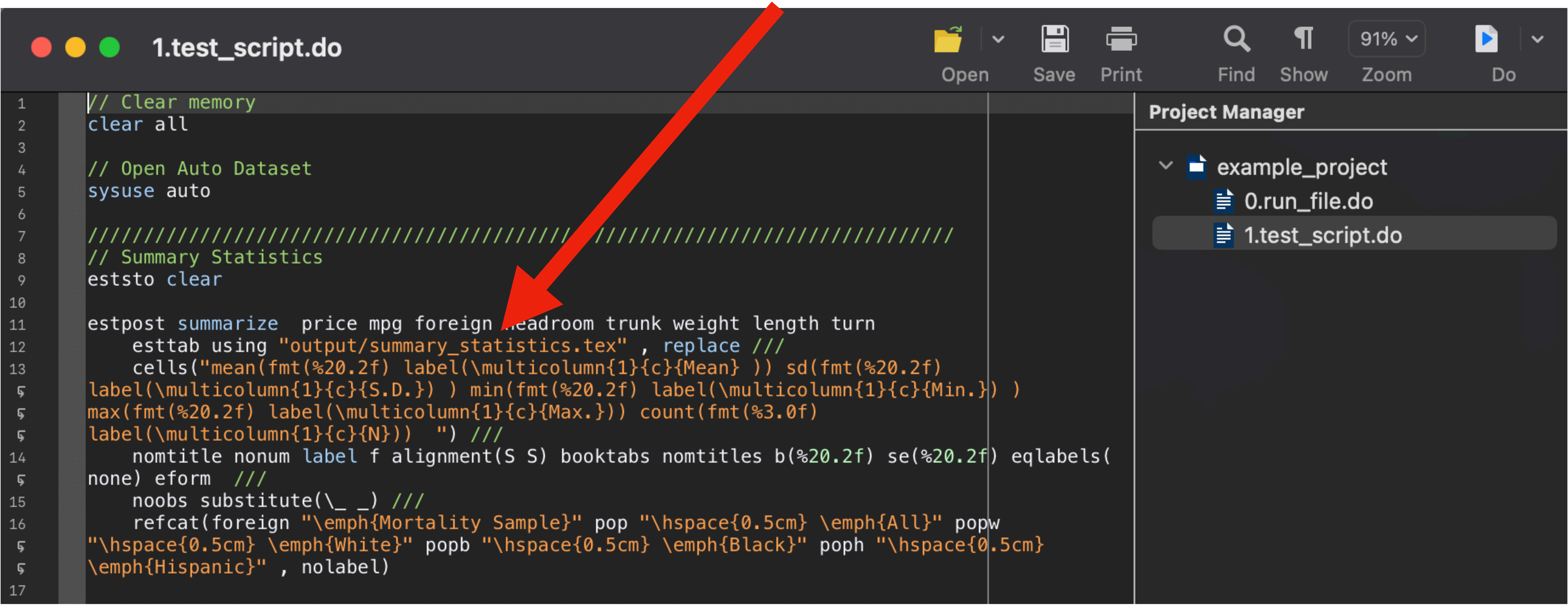
- Key benefit:
  - Can you relative file paths
  - No more hard-coding



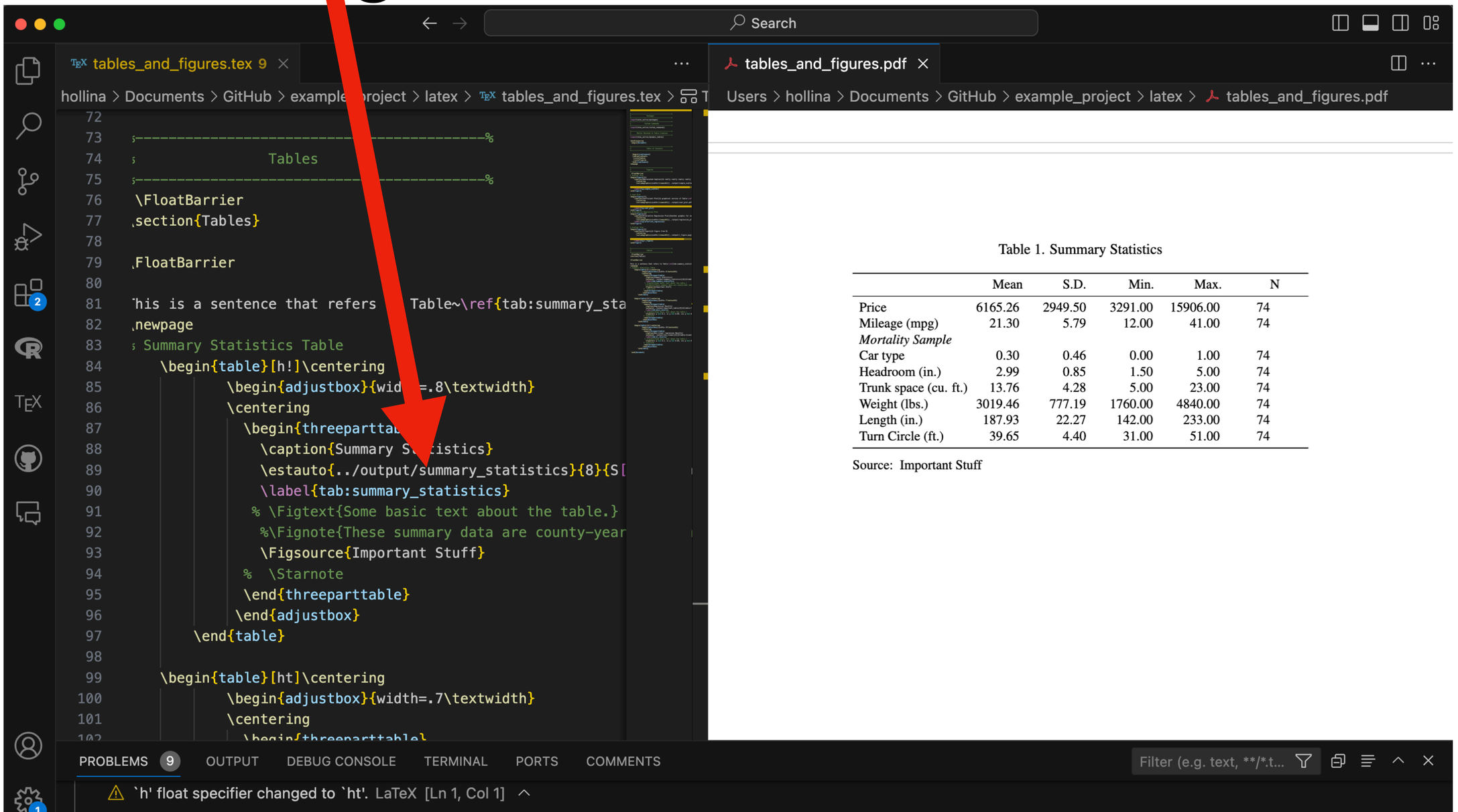
`/Users/hollina/Documents/GitHub/example_project/output/important_table.tex`

vs. `output/important_table.tex`

This really simplifies coding



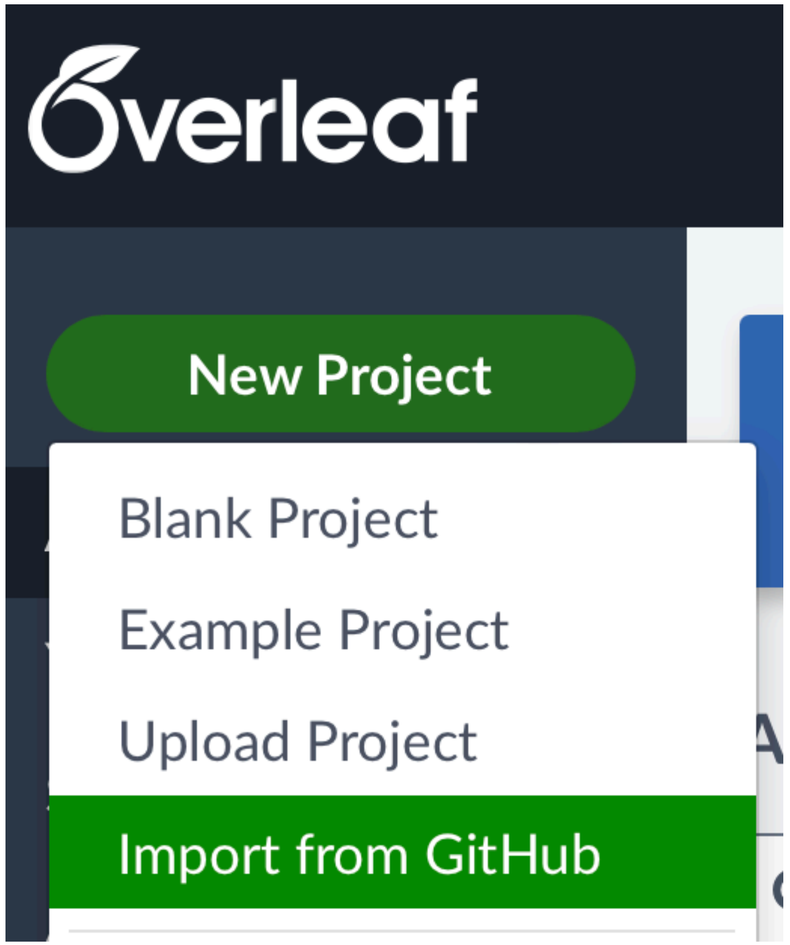
And writing


Search
tables_and_figures.tex 9
hollina > Documents > GitHub > example_project > latex > tables_and_figures.tex
tables_and_figures.pdf
Users > hollina > Documents > GitHub > example_project > latex > tables_and_figures.pdf
Tables
\FloatBarrier
\section{Tables}
\FloatBarrier
This is a sentence that refers to Table~\ref{tab:summary_sta
\newpage
% Summary Statistics Table
\begin{table}[h!]\centering
\begin{adjustbox}{wid =.8\textwidth}
\centering
\begin{threeparttable
\caption{Summary Statistics}
\estauto{../output/summary_statistics}{8}{S[
\label{tab:summary_statistics}
% \Figtext{Some basic text about the table.}
%\Fignote{These summary data are county-year
\Figsource{Important Stuff}
% \Starnote
\end{threeparttable}
\end{adjustbox}
\end{table}
\begin{table}[ht]\centering
\begin{adjustbox}{width=.7\textwidth}
\centering
\begin{threeparttable}
Table 1. Summary Statistics
Mean S.D. Min. Max. N
Price 6165.26 2949.50 3291.00 15906.00 74
Mileage (mpg) 21.30 5.79 12.00 41.00 74
Mortality Sample
Car type 0.30 0.46 0.00 1.00 74
Headroom (in.) 2.99 0.85 1.50 5.00 74
Trunk space (cu. ft.) 13.76 4.28 5.00 23.00 74
Weight (lbs.) 3019.46 777.19 1760.00 4840.00 74
Length (in.) 187.93 22.27 142.00 233.00 74
Turn Circle (ft.) 39.65 4.40 31.00 51.00 74
Source: Important Stuff
PROBLEMS 9
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS
COMMENTS
Filter (e.g. text, **/*.t...
`h' float specifier changed to `ht'. LaTeX [Ln 1, Col 1]

Speaking of writing

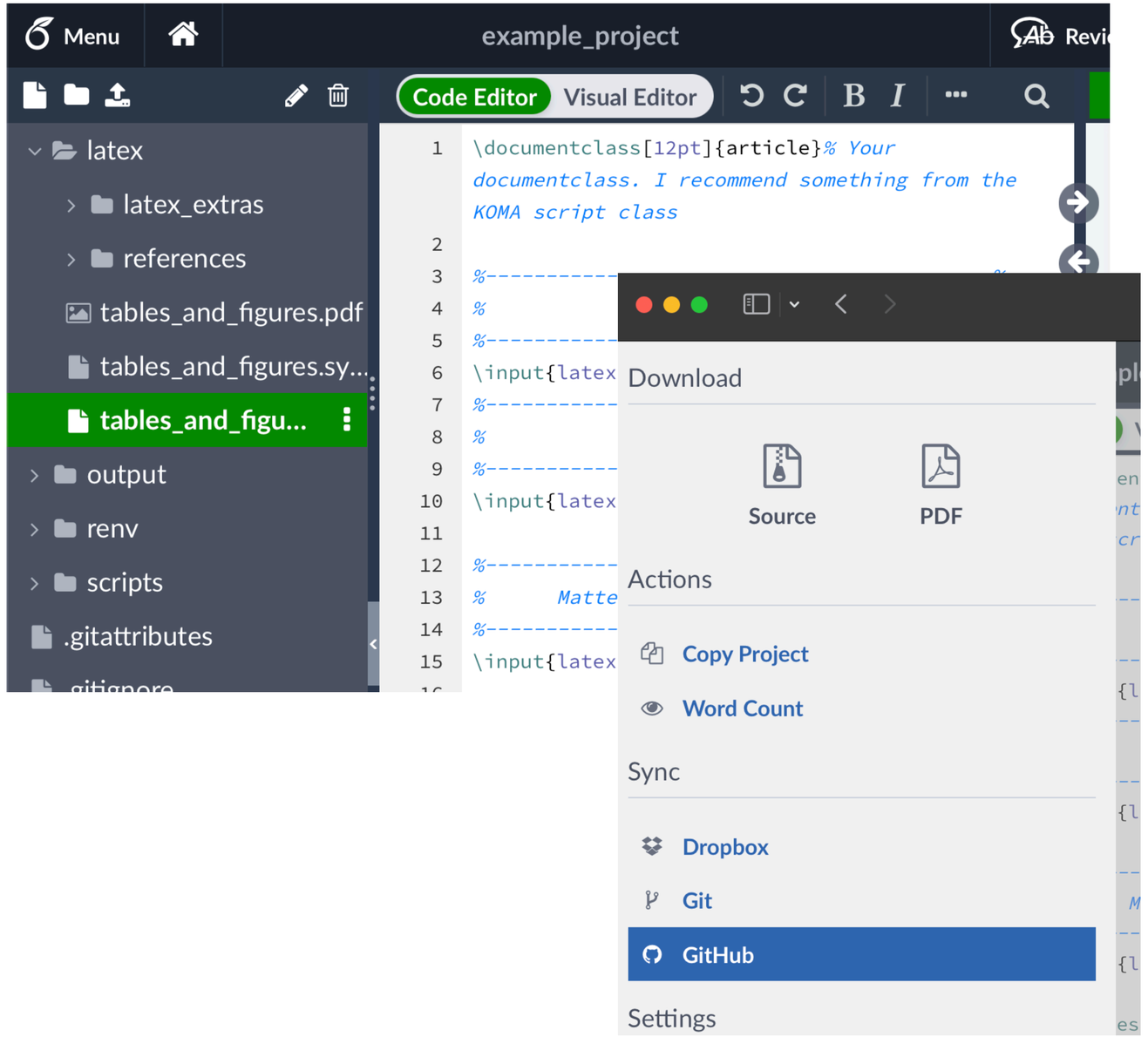
Integrating with overleaf is relatively pain free

- First, just import your GitHub project into overleaf



Integrating with overleaf is relatively pain free

- To update
  - click Menu
  - And then GitHub
  - And then write your message and click save
  - Key annoyance : It doesn't happen automatically



Packages are code too

Be sure to save the user-written packages

- Stata packages are small
  - Recommend just creating your own folder
- R packages can be large
  - renv is great
  - So is groundhog

```
37
38  ////////////////////////////////////////////////////////////
39  // Use included packages
40
41  cap adopath - PERSONAL
42  cap adopath - PLUS
43  cap adopath - SITE
44  cap adopath - OLDPLACE
45  adopath + "stata_packages"
46  net set ado "stata_packages"
47
48  // Download packages
49  if $downloads == 1 {
50      // Install  Packages
51      ssc install estout, replace
52      ssc install blindschemes, replace
53      ssc install coefplot, replace
54  }
55
56
```

Quick example

What is is?

- Download this project here: https://github.com/hollina/example_project

- Stata code
  - 0.run_file.do
    - Sets up project then runs
    - 1.test_script.do
    - 2.test_script.R
- Makes a nice latex document

Table 2. Regression Results

	(1)	(2)	(3)	(4)
Mileage (mpg)	−294.20***	−294.20***	−24.22	−24.22
	(55.69)	(60.34)	(71.71)	(79.80)
Car type	1767.29**	1767.29***	3212.15***	3212.15***
	(700.16)	(607.74)	(685.58)	(668.88)
Headroom (in.)			−625.42*	−625.42**
			(373.69)	(279.58)
Trunk space (cu. ft.)			60.94	60.94
			(92.62)	(70.66)
Weight (lbs.)			5.93***	5.93***
			(1.02)	(1.67)
Length (in.)			−74.98**	−74.98
			(37.54)	(51.16)
Turn Circle (ft.)			−157.74	−157.74
			(116.02)	(119.89)
Mean of Dependent Variable	6165.26	6165.26	6165.26	6165.26
Observations	74	74	74	74
Adjusted R sq.	0.26	0.26	0.53	0.53
Year FE	No	No	No	No
County FE	No	No	No	No
Some Type of Linear Time Trend(s)	No	No	No	No
Robust Standard Errors	No	Yes	No	Yes

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Tools/Resources

- GitHub: https://github.com/
- GitHub Desktop: https://desktop.github.com/
- VSCode: https://code.visualstudio.com/
- GitHub CoPilot: https://github.com/features/copilot

- Julian Reif’s coding guide: https://julianreif.com/guide/