10 Steps to Becoming a Tidyverse Contributor

Nic Crane

Hello!

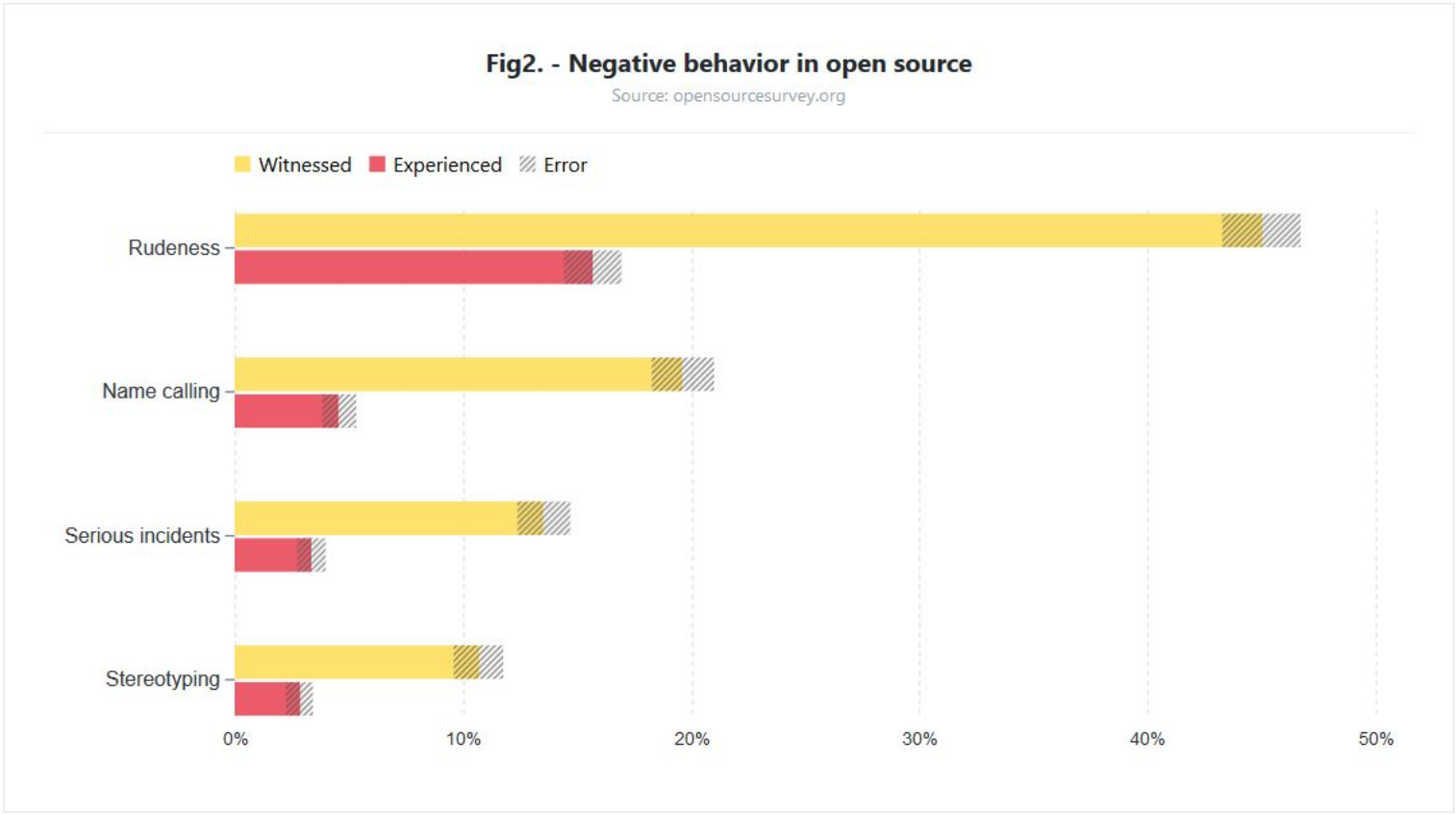
Nic Crane

@nic_crane

Who here…

- Knows what the Tidyverse is?
- Tweets about R?
- Blogs about R?
- Has written an R function?
- Uses git?
- Has written an R package?
- Has a package on CRAN?
- Has contributed to a Tidyverse package?

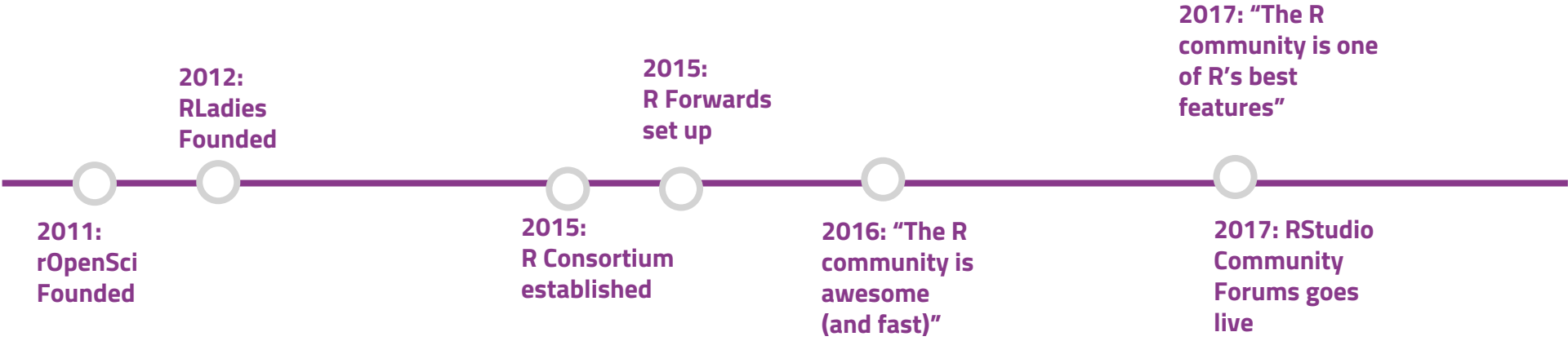
https://opensourcesurvey.org/2017/#insights



Culture is Changing
in Open Source

stackoverflow

How Does R Compare?



Why Should You Contribute?

GITHUB SURVEY 2017

- “Half of contributors say that their open source work was somewhat or very important in getting their current role.”
- “Open source work helps people build their professional reputation.”
- “Improving contributor representation can help create a more representative tech sector overall.”

Why Should You Contribute?

MY PERSONAL REASONS

- Improve code knowledge
- Interact with the community
- ~~Overcome~~ Lessen imposter syndrome

When I started my transition into data science, I said yes to pretty much every opportunity that came my way, even if it felt slightly beyond my skill set or experience level...I said yes to many of these things when it felt like I wasn't sure if I was ready.

—*Julia Silge*

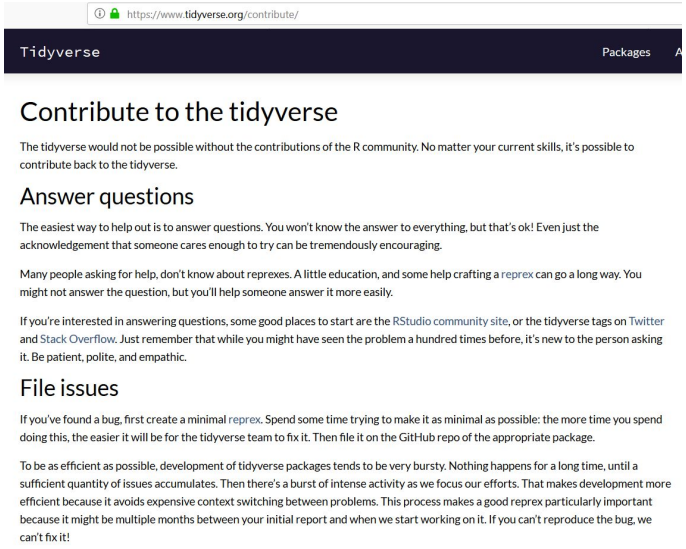
Interview with And Comfort (October 2018)

Not always easy for beginners. Helps to have a team that will onboard. Give beginner level work. Mentorship. Pick projects that are welcoming.

—Gabriela de Queiroz [summarised tweet]

Grace Hopper Celebration of Women in Computing (2018)

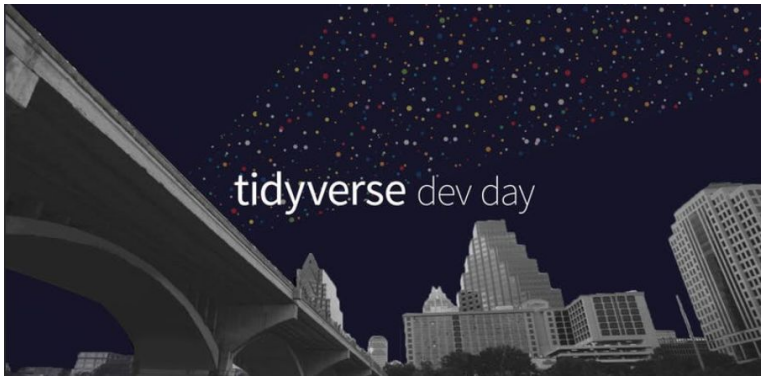
The tidyverse is a great place to get involved



beginner-friendly

help wanted

good first issue



Step 1
Decide How to Contribute

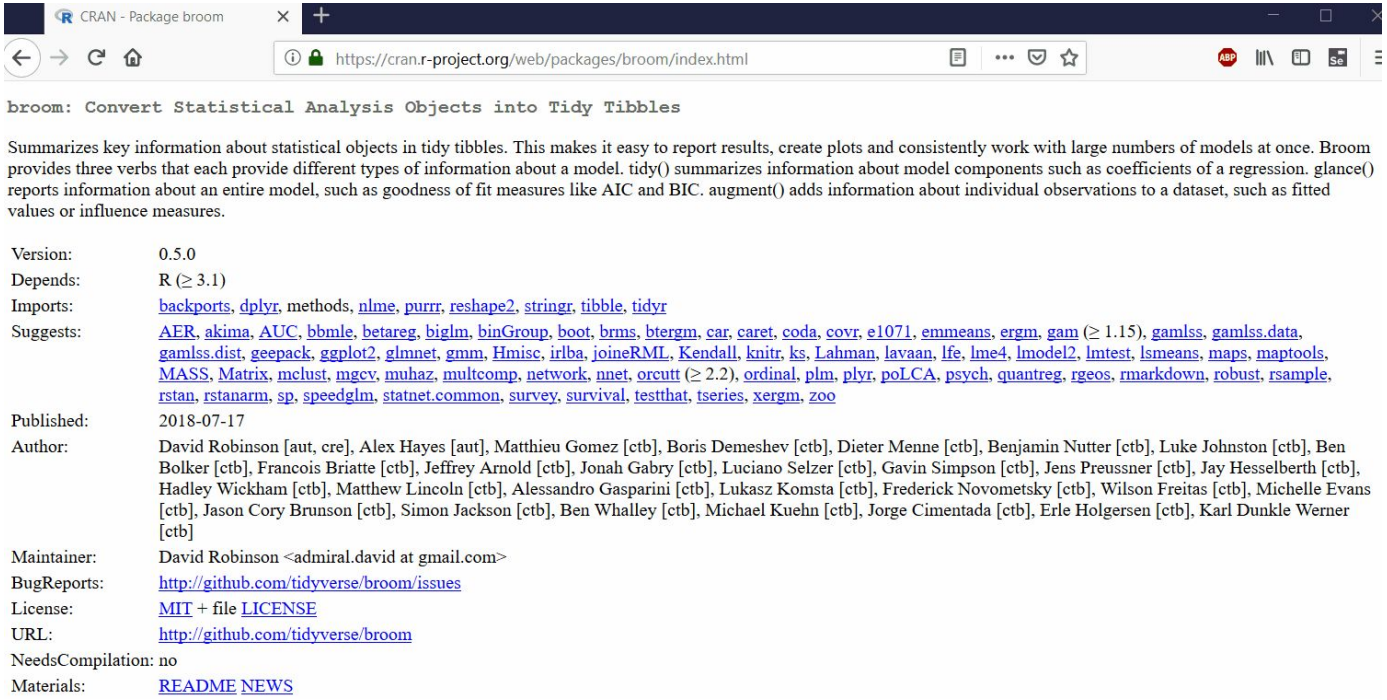
- Identifying issues
- **Documentation**
- Helping people out on Twitter and forums
- Blogging
- **Code**

Step 2
Learn!

- R functions
- git (clone, commit, push)
- Building R packages (basics)
- GitHub – making a pull request

Step 3
Find an Issue

Step 3
Find an Issue

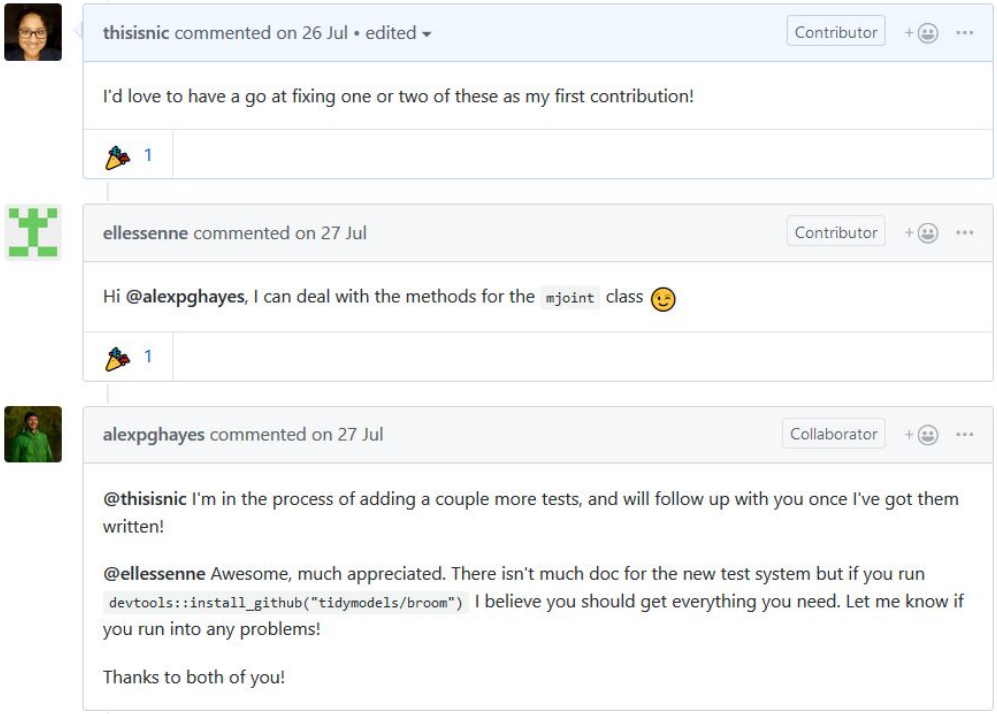


broom: Convert Statistical Analysis Objects into Tidy Tibbles

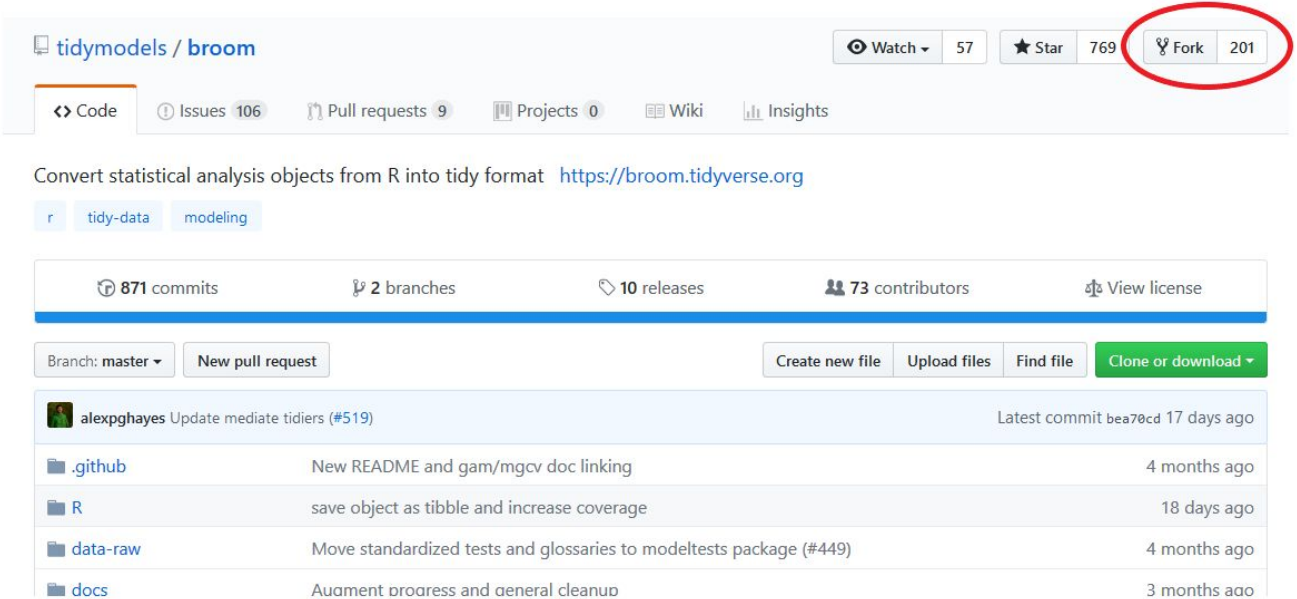
Summarizes key information about statistical objects in tidy tibbles. This makes it easy to report results, create plots and consistently work with large numbers of models at once. Broom provides three verbs that each provide different types of information about a model. tidy() summarizes information about model components such as coefficients of a regression. glance() reports information about an entire model, such as goodness of fit measures like AIC and BIC. augment() adds information about individual observations to a dataset, such as fitted values or influence measures.

Version:	0.5.0
Depends:	R (≥ 3.1)
Imports:	backports, dplyr, methods, nlme, purrr, reshape2, stringr, tibble, tidyr
Suggests:	AER, akima, AUC, bbmle, betareg, biglm, binGroup, boot, brms, btergm, car, caret, coda, covr, e1071, emmeans, ergm, gam (≥ 1.15), gamlss, gamlss.data, gamlss.dist, geepack, ggplot2, glmnet, gmm, Hmisc, irlba, joineRML, Kendall, knitr, ks, Lahman, lavaan, lfe, lme4, lmodel2, lmtest, lsmeans, maps, maptools, MASS, Matrix, mclust, mgcv, muhaz, multcomp, network, nnet, orcutt (≥ 2.2), ordinal, plm, plyr, poLCA, psych, quantreg, rgeos, rmarkdown, robust, rsample, rstan, rstanarm, sp, speedglm, statnet.common, survey, survival, testthat, tseries, xergm, zoo
Published:	2018-07-17
Author:	David Robinson [aut, cre], Alex Hayes [aut], Matthieu Gomez [ctb], Boris Demeshev [ctb], Dieter Menne [ctb], Benjamin Nutter [ctb], Luke Johnston [ctb], Ben Bolker [ctb], Francois Briatte [ctb], Jeffrey Arnold [ctb], Jonah Gabry [ctb], Luciano Selzer [ctb], Gavin Simpson [ctb], Jens Preussner [ctb], Jay Hesselberth [ctb], Hadley Wickham [ctb], Matthew Lincoln [ctb], Alessandro Gasparini [ctb], Lukasz Komsta [ctb], Frederick Novometsky [ctb], Wilson Freitas [ctb], Michelle Evans [ctb], Jason Cory Brunson [ctb], Simon Jackson [ctb], Ben Whalley [ctb], Michael Kuehn [ctb], Jorge Cimentada [ctb], Erle Holgersen [ctb], Karl Dunkle Werner [ctb]
Maintainer:	David Robinson <admiral.david at gmail.com>
BugReports:	http://github.com/tidyverse/broom/issues
License:	MIT + file LICENSE
URL:	http://github.com/tidyverse/broom
NeedsCompilation:	no
Materials:	README NEWS

Step 4
Ask if You Can Help

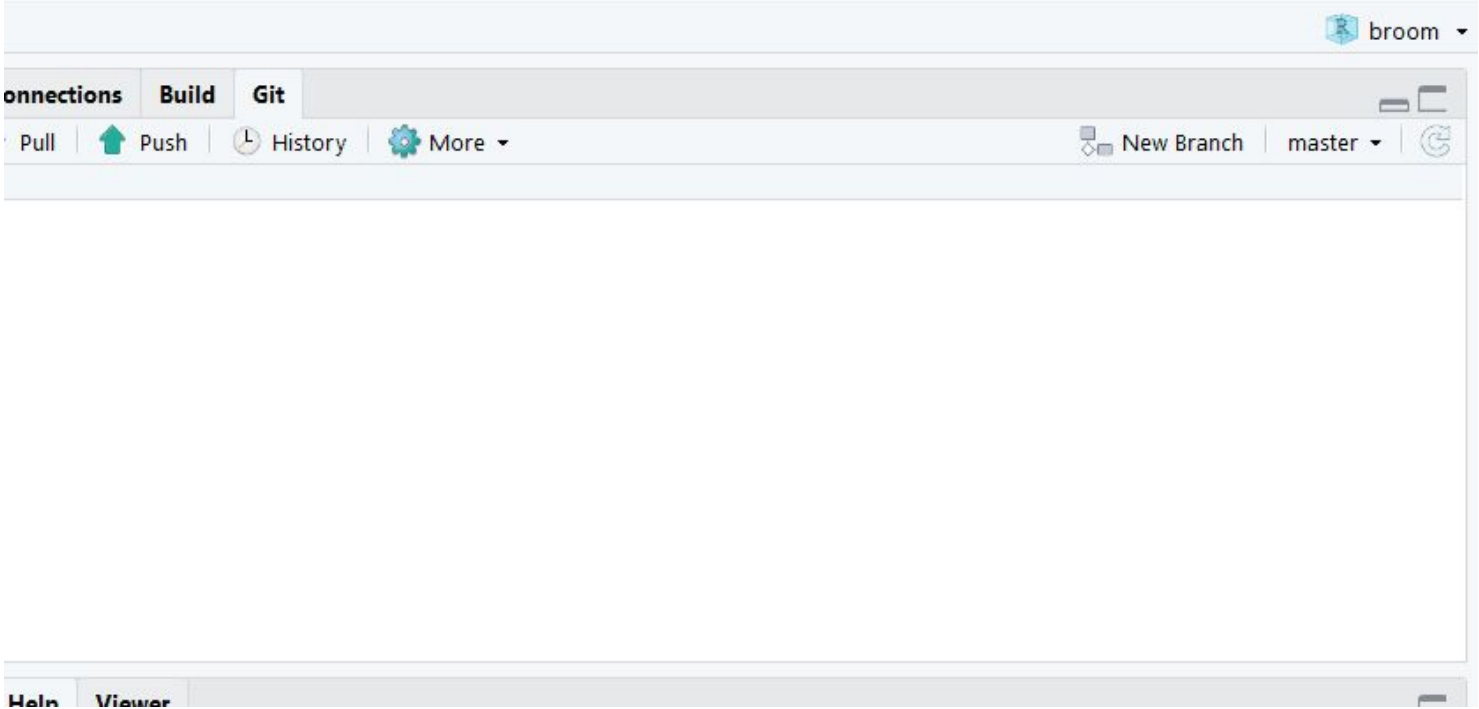


Step 5
Fork the Repo

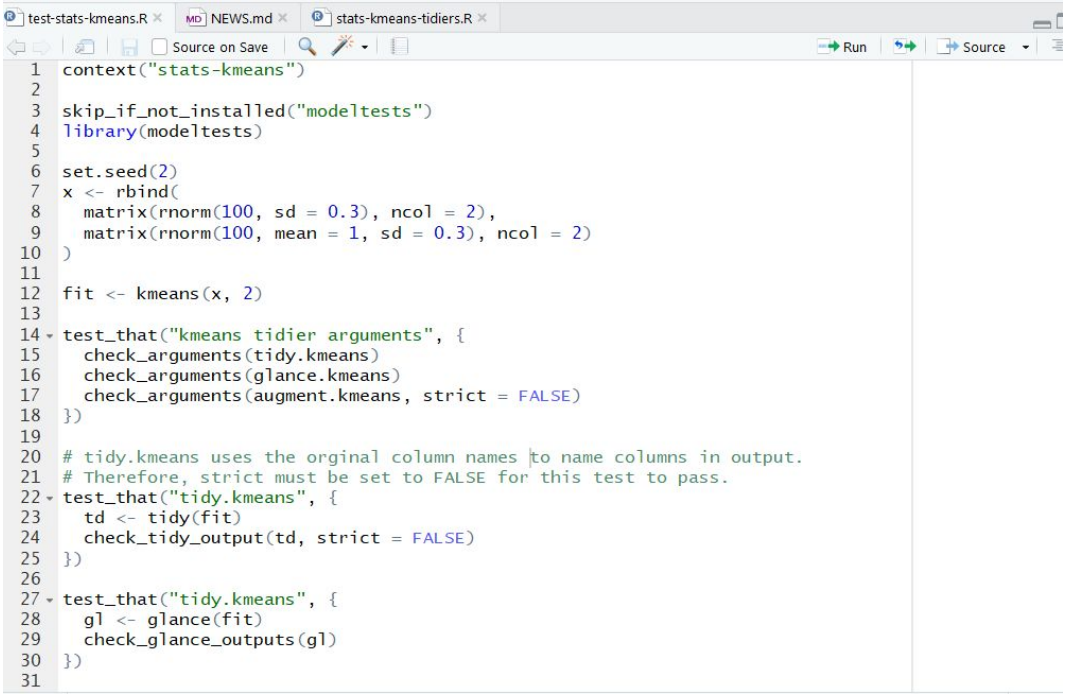


Step 6

Open in RStudio & Make a Branch



Step 7
Familiarise Yourself with the Code

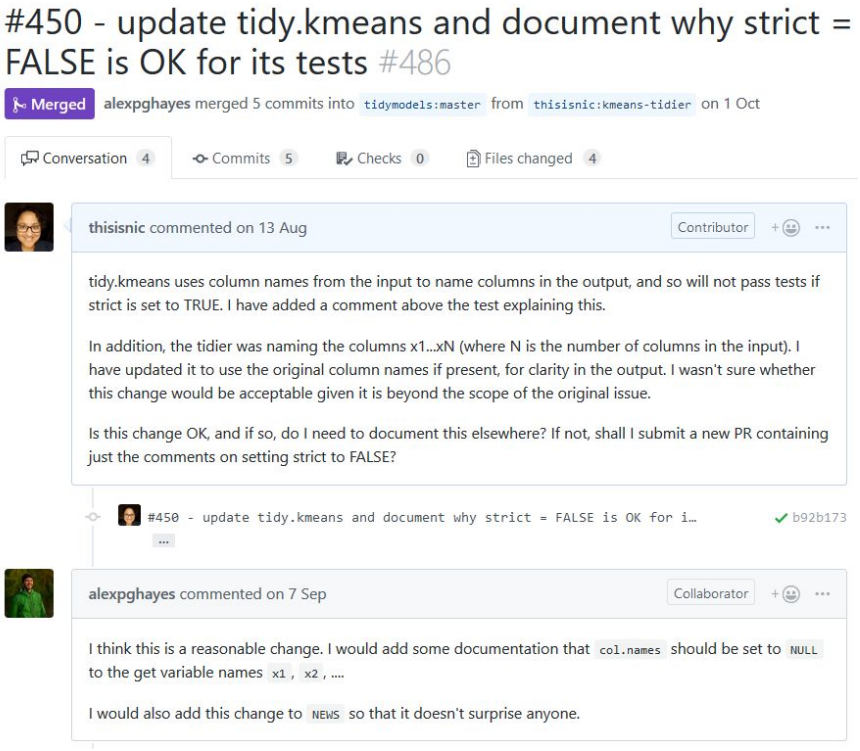


Step 8
Write Code

8 R/stats-kmeans-tidiers.R — View

```
@@ -2,7 +2,7 @@
 #' @template title_desc_tidy
 #'
 #' @param x A `kmeans` object created by [stats::kmeans()].
-#' @param col.names Dimension names. Defaults to `x1, x2, ...`
+#' @param col.names Dimension names. Defaults to the names of the variables in x.  Set to NULL to get names `x1, x2, ...`.
 #' @template param_unused_dots
 #'
 #' @evalRd return_tidy("size", "withinss", "cluster")
@@ -13,7 +13,11 @@
 #' @export
 #' @seealso [tidy()], [stats::kmeans()]
 #' @family kmeans tidiers
-tidy.kmeans <- function(x, col.names = paste0("x", 1:ncol(x$centers)), ...) {
+tidy.kmeans <- function(x, col.names = colnames(x$centers), ...) {
+
+  if(is.null(col.names)){
+    col.names <- paste0("x", 1:ncol(x$centers))
+  }
   ret <- as.data.frame(x$centers)
   colnames(ret) <- col.names
   ret$size <- x$size
```

Step 9
Submit PR and Wait



#450 - update tidy.kmeans and document why strict = FALSE is OK for its tests #486

Merged **alexpghayes** merged 5 commits into `tidymodels:master` from `thisisnic:kmeans-tidier` on 1 Oct

Conversation 4 · Commits 5 · Checks 0 · Files changed 4

thisisnic commented on 13 Aug — Contributor

tidy.kmeans uses column names from the input to name columns in the output, and so will not pass tests if strict is set to TRUE. I have added a comment above the test explaining this.

In addition, the tidier was naming the columns x1...xN (where N is the number of columns in the input). I have updated it to use the original column names if present, for clarity in the output. I wasn't sure whether this change would be acceptable given it is beyond the scope of the original issue.

Is this change OK, and if so, do I need to document this elsewhere? If not, shall I submit a new PR containing just the comments on setting strict to FALSE?

`#450 - update tidy.kmeans and document why strict = FALSE is OK for i…` ✓ b92b173

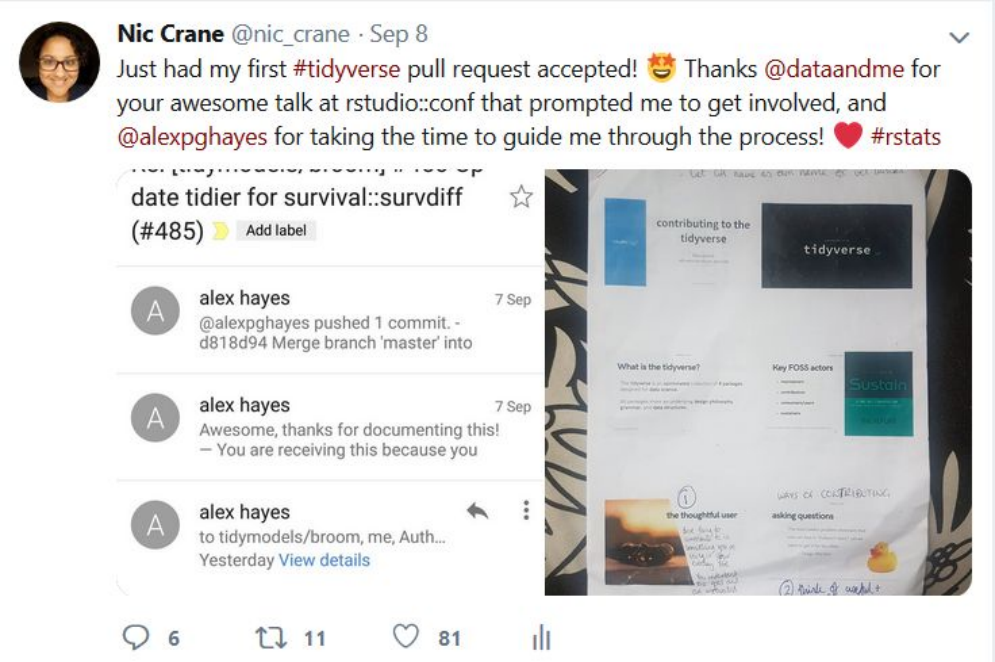
alexpghayes commented on 7 Sep — Collaborator

I think this is a reasonable change. I would add some documentation that `col.names` should be set to `NULL` to the get variable names `x1`, `x2`, ...

I would also add this change to `NEWS` so that it doesn't surprise anyone.

Step 10

Celebrate! Then Encourage Others to Get Involved and Share Your Experience



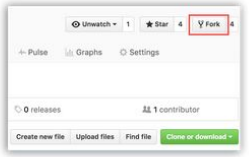
Never Submitted a PR Before?

Check out the first contributions repo!

https://github.com/thisisnic/first-contributions

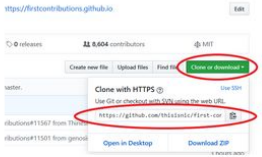
Fork this repository

Fork this repo by clicking on the fork button on the top of this page.



This will create of copy of this repository in your account and you will be redirected to this version of the repo.

You now need to get the URL for your repo. Click "Clone or download" and copy the URL in the box.



Create a new project

In RStudio, go to File -> New Project

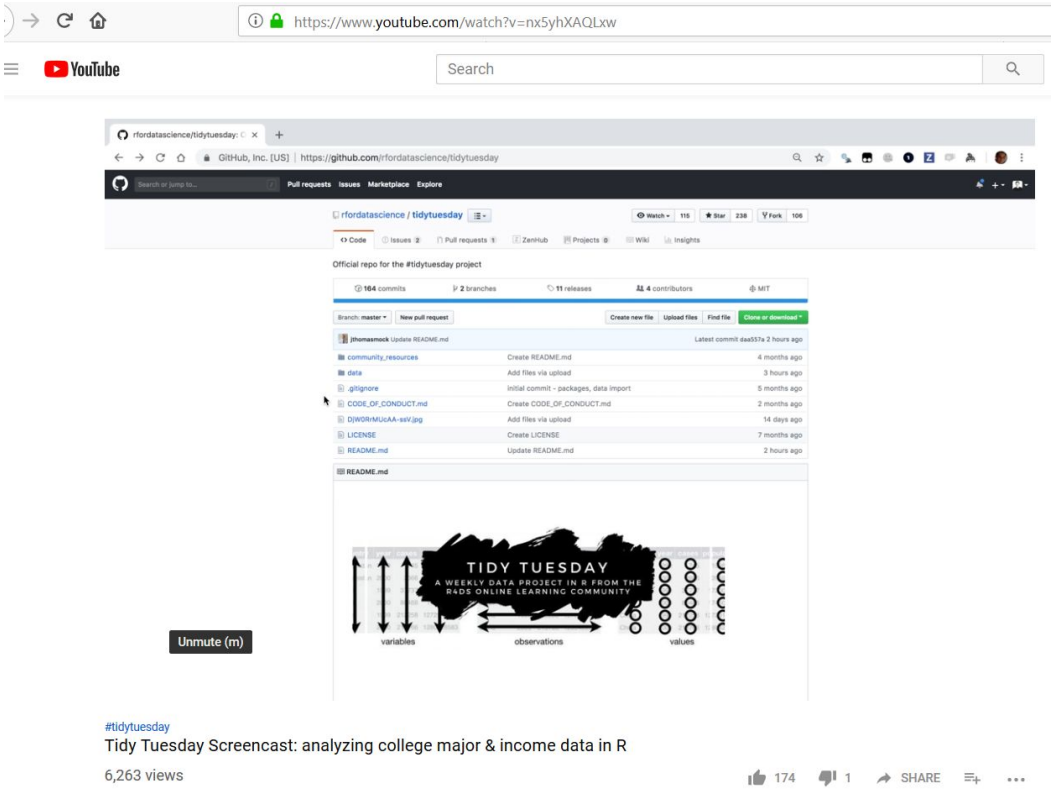
Other ways to get involved

- Blogging
- Screencasts
- Tweets
- GitHub gists

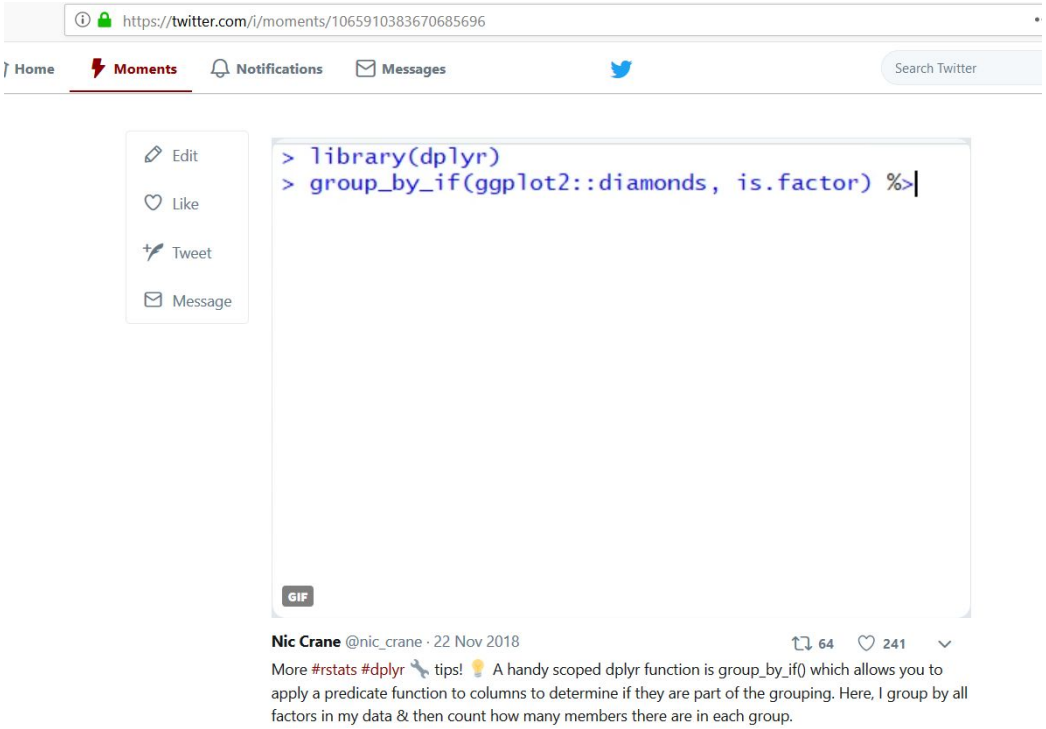
Blogging


https://suzan.rbind.io/2018/01/dplyr-tutorial-1/
Suzan Baert
Home
About me
Tutorials
Archives
GitHub
Twitter
RSS
Data Wrangling Part 1: Basic to Advanced Ways to Select Columns
January 31, 2018 in Tutorial
I went through the entire dplyr documentation for a talk last week about pipes, which resulted in a few "aha!" moments. I discovered and re-discovered a few useful functions, which I wanted to collect in a few blog posts so I can share them with others.
This first post will cover ordering, naming and selecting columns, it covers the basics of selecting columns and more advanced functions like select_all() , select_if() and shortcuts like everything() .
Other posts in this series:
Part 2: Transforming your columns into the right shape, which covers functions such as mutate_all() , mutate_if() and mutate_at() , but also ways to code into discrete columns with case_when() .
Part 3: Filtering rows, which includes tricks like the nifty near() functions and the filter_all / fitler_if / filter_at family.
Part 4: Summarising your data, which includes the add_count() shortcut, and the summarise_all / summarise_if / summarise_at family.

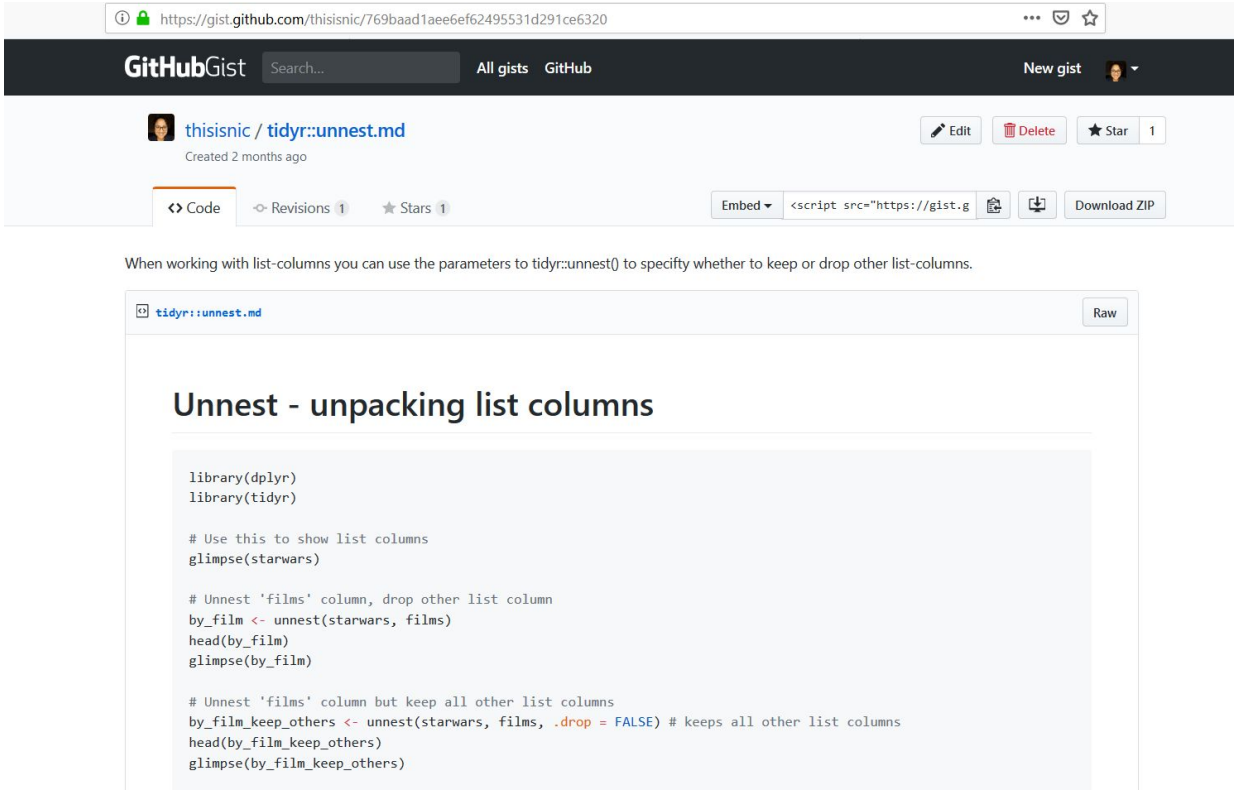
Screencasts



Tweeting

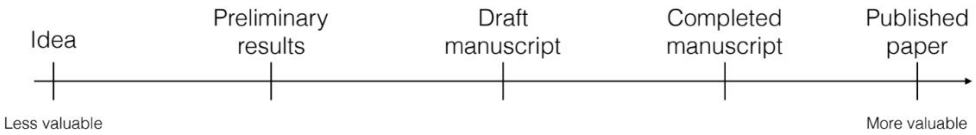

https://twitter.com/i/moments/1065910383670685696
Home
Moments
Notifications
Messages
Search Twitter
Edit
Like
Tweet
Message
> library(dplyr)
> group_by_if(ggplot2::diamonds, is.factor) %>
GIF
Nic Crane @nic_crane · 22 Nov 2018
64
241
More #rstats #dplyr 🔧 tips! 💡 A handy scoped dplyr function is group_by_if() which allows you to apply a predicate function to columns to determine if they are part of the grouping. Here, I group by all factors in my data & then count how many members there are in each group.

GitHub Gists



“The unreasonable effectiveness of public work” – David Robinson

How I used to think of my goals:



How I should have been thinking of them:

Resources

Talk by Mara Averick on contributing:
https://www.rstudio.com/resources/videos/contributing-to-tidyverse-packages/

rOpenSci contributing guide:
https://ropensci.github.io/dev_guide/contributingguide.html

Repo walking through making a pull request:
https://github.com/thisisnic/first-contributions

Guide to contributing code to the tidyverse:
https://www.tidyverse.org/articles/2017/08/contributing/

Talk by David Robinson on public work:
https://resources.rstudio.com/rstudio-conf-2019/the-unreasonable-effectiveness-of-public-work

Thanks! ☺

Nic Crane
@nic_crane