



| 10 Steps to Becoming a Tidyverse Contributor

Nic Crane



Hello!

Nic Crane

@nic_crane 

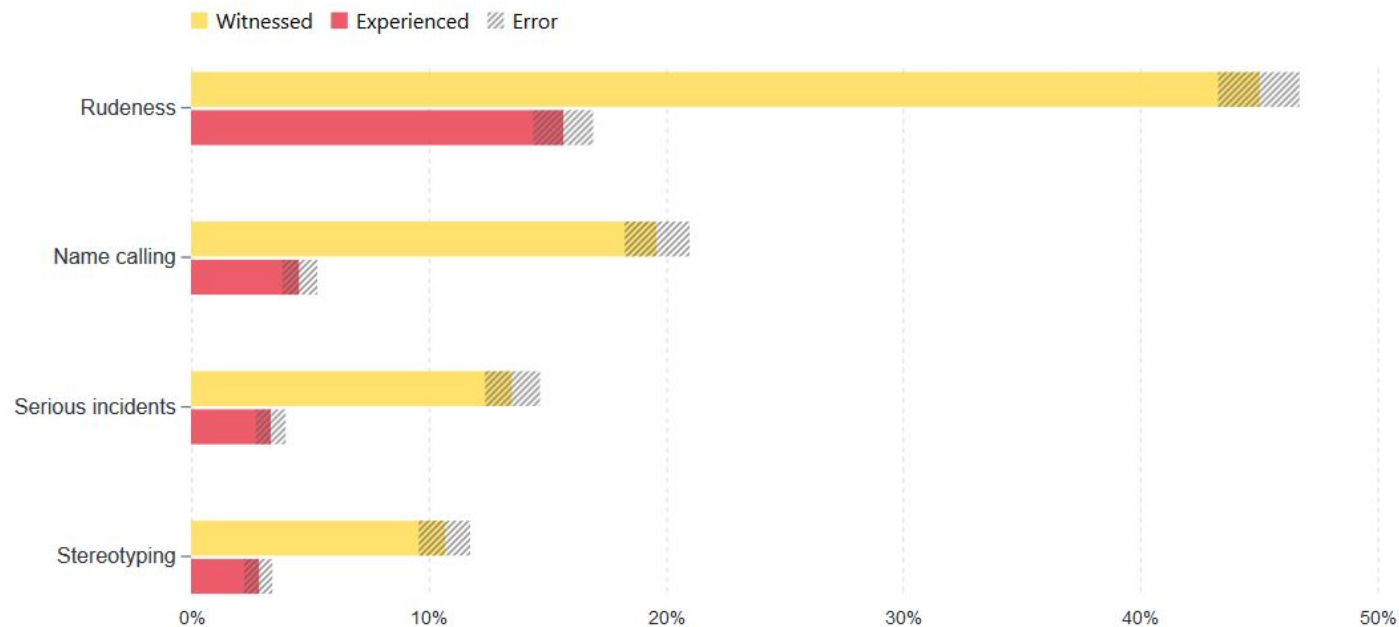


Who here...

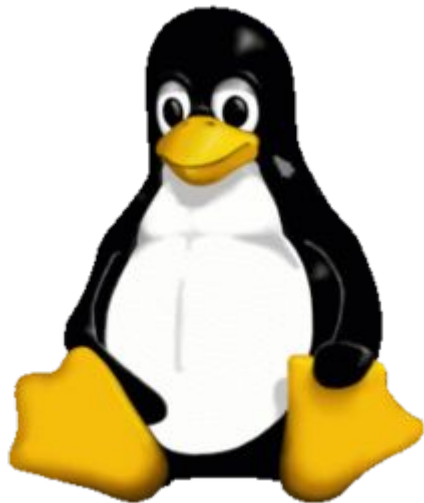
- **Knows what the Tidyverse is?**
- **Tweets about R?**
- **Blogs about R?**
- **Has written an R function?**
- **Uses git?**
- **Has written an R package?**
- **Has a package on CRAN?**
- **Has contributed to a Tidyverse package?**

Fig2. - Negative behavior in open source

Source: opensourcesurvey.org



Culture is Changing in Open Source



How Does R Compare?



2012:
RLadies
Founded

2015:
R Forwards
set up

2017: "The R
community is one
of R's best
features"

2011:
rOpenSci
Founded

2015:
R Consortium
established

2016: "The R
community is
awesome
(and fast)"

2017: RStudio
Community
Forums goes
live



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)

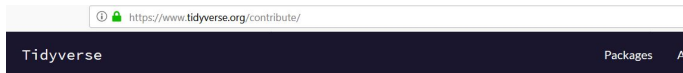
The tidyverse is a great place to get involved



beginner-friendly

help wanted

good first issue



Contribute to the tidyverse

The tidyverse would not be possible without the contributions of the R community. No matter your current skills, it's possible to contribute back to the tidyverse.

Answer questions

The easiest way to help out is to answer questions. You won't know the answer to everything, but that's ok! Even just the acknowledgement that someone cares enough to try can be tremendously encouraging.

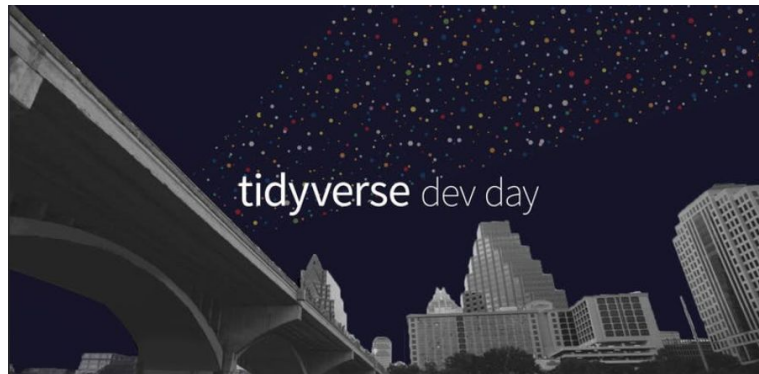
Many people asking for help, don't know about reprexes. A little education, and some help crafting a reprex can go a long way. You might not answer the question, but you'll help someone answer it more easily.

If you're interested in answering questions, some good places to start are the RStudio community site, or the tidyverse tags on Twitter and Stack Overflow. Just remember that while you might have seen the problem a hundred times before, it's new to the person asking it. Be patient, polite, and empathic.

File issues

If you've found a bug, first create a minimal reprex. Spend some time trying to make it as minimal as possible: the more time you spend doing this, the easier it will be for the tidyverse team to fix it. Then file it on the GitHub repo of the appropriate package.

To be as efficient as possible, development of tidyverse packages tends to be very bursty. Nothing happens for a long time, until a sufficient quantity of issues accumulates. Then there's a burst of intense activity as we focus our efforts. That makes development more efficient because it avoids expensive context switching between problems. This process makes a good reprex particularly important because it might be multiple months between your initial report and when we start working on it. If you can't reproduce the bug, we can't fix it!





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

beginner-friendly

help wanted

good first issue

Step 3

Find an Issue



CRAN - Package broom

https://cran.r-project.org/web/packages/broom/index.html

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](#), [lmtree](#), [lsmeans](#), [maps](#), [maptools](#), [MASS](#), [Matrix](#), [mclust](#), [mgcv](#), [muhaz](#), [multcomp](#), [network](#), [nnet](#), [oreut](#) (≥ 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@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



thisisnic commented on 26 Jul • edited ▼

Contributor



I'd love to have a go at fixing one or two of these as my first contribution!



1



ellessenne commented on 27 Jul

Contributor



Hi @alexpghayes, I can deal with the methods for the `mjoint` class 😊



1



alexpghayes commented on 27 Jul

Collaborator



@thisisnic I'm in the process of adding a couple more tests, and will follow up with you once I've got them written!

@ellessenne Awesome, much appreciated. There isn't much doc for the new test system but if you run `devtools::install_github("tidymodels/broom")` I believe you should get everything you need. Let me know if you run into any problems!

Thanks to both of you!



Step 5

Fork the Repo

tidymodels / broom

Watch 57 Star 769 Fork 201

Code Issues 106 Pull requests 9 Projects 0 Wiki Insights

Convert statistical analysis objects from R into tidy format <https://broom.tidyverse.org>

r tidy-data modeling

871 commits 2 branches 10 releases 73 contributors View license

Branch: master New pull request Create new file Upload files Find file Clone or download

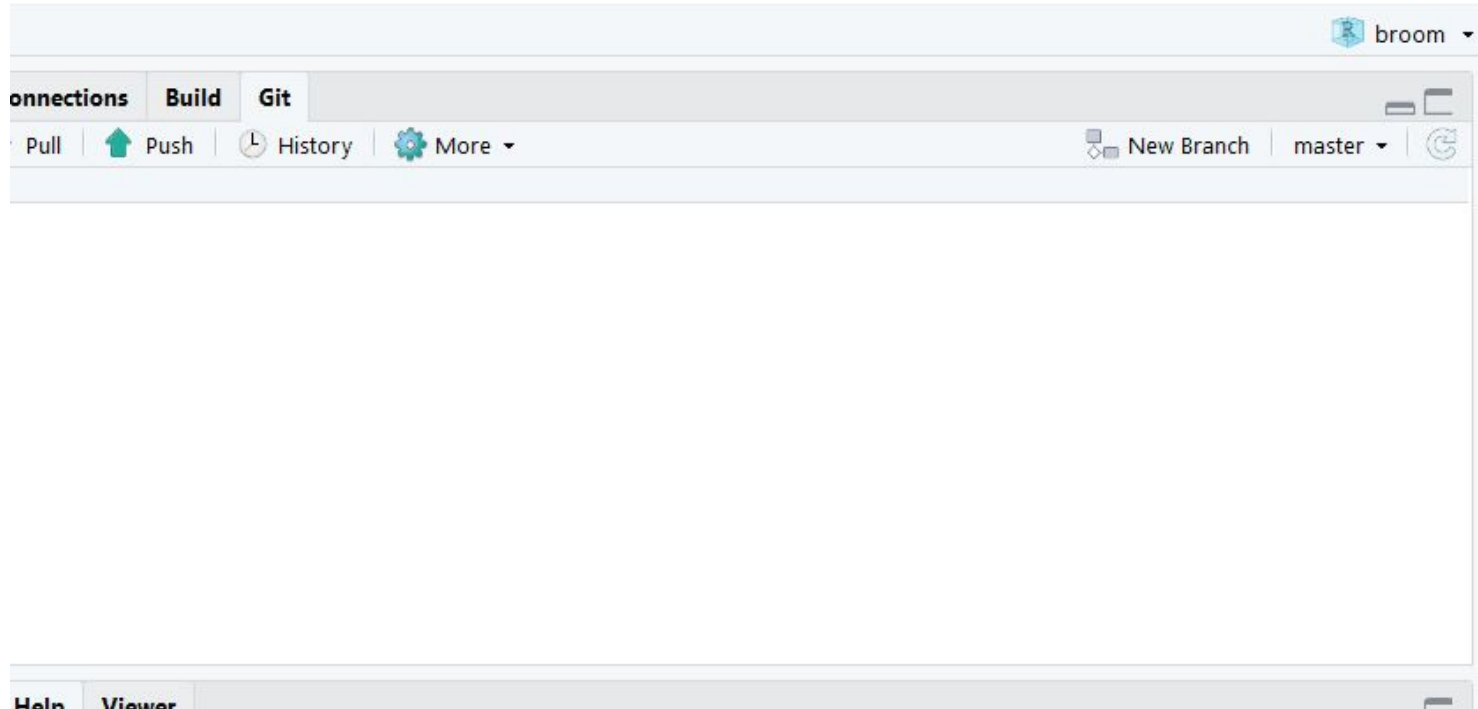
alexphayes Update mediate tidiers (#519) Latest commit bea70cd 17 days ago

.github	New README and gam/mgcv doc linking	4 months ago
R	save object as tibble and increase coverage	18 days ago
data-raw	Move standardized tests and glossaries to modeltests package (#449)	4 months ago
docs	Auament broaress and aeneral cleanup	3 months ago



Step 6

Open in RStudio & Make a Branch





Step 7

Familiarise Yourself with the Code

```
test-stats-kmeans.R | MD NEWS.md | stats-kmeans-tidiers.R
Source on Save
Run Source

1 context("stats-kmeans")
2
3 skip_if_not_installed("modeltests")
4 library(modeltests)
5
6 set.seed(2)
7 x <- rbind(
8   matrix(rnorm(100, sd = 0.3), ncol = 2),
9   matrix(rnorm(100, mean = 1, sd = 0.3), ncol = 2)
10 )
11
12 fit <- kmeans(x, 2)
13
14 test_that("kmeans tidier arguments", {
15   check_arguments(tidy.kmeans)
16   check_arguments(glance.kmeans)
17   check_arguments(augment.kmeans, strict = FALSE)
18 })
19
20 # tidy.kmeans uses the original column names to name columns in output.
21 # Therefore, strict must be set to FALSE for this test to pass.
22 test_that("tidy.kmeans", {
23   td <- tidy(fit)
24   check_tidy_output(td, strict = FALSE)
25 })
26
27 test_that("tidy.kmeans", {
28   gl <- glance(fit)
29   check_glance_outputs(gl)
30 })
31
```

Step 8

Write Code



8 R/stats-kmeans-tidiers.R

View



@@ -2,7 +2,7 @@

```
2 #' @template title_desc_tidy
3 #'
4 #' @param x A `kmeans` object created by [stats::kmeans()].
5 - #' @param col.names Dimension names. Defaults to `x1, x2, ...`
6
7 #' @template param_unused_dots
8 #'
9 #' @evalRd return_tidy("size", "withinss", "cluster")
```

@@ -13,7 +13,11 @@

```
13 #' @export
14 #' @seealso [tidy()], [stats::kmeans()]
15 #' @family kmeans tidiers
16 - tidy.kmeans <- function(x, col.names = paste0("x", 1:ncol(x$centers)), ...) {
17
18   ret <- as.data.frame(x$centers)
19   colnames(ret) <- col.names
20   ret$size <- x$size
21 }
```

```
2 #' @template title_desc_tidy
3 #'
4 #' @param x A `kmeans` object created by [stats::kmeans()].
5 + #' @param col.names Dimension names. Defaults to the names of the variables in x. Set to NULL to get
6   names `x1, x2, ...`.
7
8 #' @template param_unused_dots
9 #'
10 #' @evalRd return_tidy("size", "withinss", "cluster")
11
12
13 #' @export
14 #' @seealso [tidy()], [stats::kmeans()]
15 #' @family kmeans tidiers
16 + tidy.kmeans <- function(x, col.names = colnames(x$centers), ...) {
17 +
18 +   if(is.null(col.names)){
19 +     col.names <- paste0("x", 1:ncol(x$centers))
20 +   }
21
22   ret <- as.data.frame(x$centers)
23   colnames(ret) <- col.names
24   ret$size <- x$size
25 }
```



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

Nic Crane @nic_crane · Sep 8
Just had my first #tidyverse pull request accepted! 🥳 Thanks @dataandme for your awesome talk at rstudio::conf that prompted me to get involved, and @alexpghayes for taking the time to guide me through the process! ❤️ #rstats

for tidy models, broom, me, Auth...
date tidier for survival::survdiff
(#485) Add label

alex hayes @alexpghayes pushed 1 commit. - d818d94 Merge branch 'master' into
7 Sep

alex hayes 7 Sep
Awesome, thanks for documenting this!
— You are receiving this because you

alex hayes to tidymodels/broom, me, Auth... Yesterday View details

6 11 81

The diagram is a hand-drawn poster titled 'contributing to the tidyverse' and 'tidyverse'. It outlines the process of contributing to the tidyverse ecosystem. It includes sections for 'What is the tidyverse?', 'Key FOSS actors', 'Sustain', 'the thoughtful user', and 'asking questions'. It also features a small illustration of a person sitting at a desk with a laptop and a rubber duck.



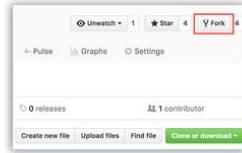
Never Submitted a PR Before?

Check out the first contributions repo!

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

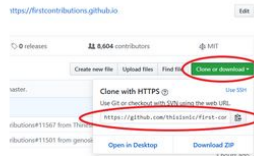
Fork this repository

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



This will create a 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



The screenshot shows a web browser displaying a blog post. The browser's address bar shows the URL `https://suzan.rbind.io/2018/01/dplyr-tutorial-1/`. The blog's sidebar on the left features a profile picture of Suzan Baert and a list of navigation links: Home, About me, Tutorials, Archives, GitHub, Twitter, and RSS. The main content area displays the article title "Data Wrangling Part 1: Basic to Advanced Ways to Select Columns" with a sub-header "January 31, 2018 in Tutorial". The text of the post begins with "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." It then states, "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()`." Below this, a section titled "Other posts in this series:" lists three links: "Part 2: Transforming your columns into the right shape", "Part 3: Filtering rows", and "Part 4: Summarising your data". Each link is followed by a brief description of the topics it covers, including specific R functions like `mutate_all()`, `mutate_if()`, `mutate_at()`, `near()`, `filter_all()`, `filter_if()`, `filter_at()`, `add_count()`, `summarise_all()`, `summarise_if()`, and `summarise_at()`.

← → ↺ 🏠 `https://suzan.rbind.io/2018/01/dplyr-tutorial-1/` 📄 ⋮ 📖 ☆

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` / `filter_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



→ ↺ 🏠 <https://www.youtube.com/watch?v=nx5yhXaQLxw>

☰ YouTube Search 🔍

rfordata science / tidyuesday: 1 x +

← → 🔄 🌐 GitHub, Inc. (US) | <https://github.com/rfordata science/tidyuesday> 🔍 ⭐ 238 🍴 106

🔍 Search or jump to... Pull requests Issues Marketplace Explore

rfordata science / tidyuesday

🔗 Code 📄 Issues 3 📄 Pull requests 3 📄 Zenhub 📄 Projects 0 📄 Wiki 📄 Insights

Official repo for the #tidyuesday project

📄 164 commits 📄 2 branches 📄 11 releases 📄 4 contributors 📄 MIT

Branch: master ← New pull request Create new file Upload file Find file Clone or download

📄 shomasek Update README.md	Latest commit d4d537a 2 hours ago
📄 community_resources	Create README.md 4 months ago
📄 data	Add files via upload 3 hours ago
📄 .gitignore	Initial commit - packages, data import 5 months ago
📄 CODE_OF_CONDUCT.md	Create CODE_OF_CONDUCT.md 2 months ago
📄 DOWRMUGAA-wr.png	Add files via upload 14 days ago
📄 LICENSE	Create LICENSE 7 months ago
📄 README.md	Update README.md 2 hours ago

📄 README.md

The logo for Tidy Tuesday, featuring the text "TIDY TUESDAY" in a bold, black, hand-drawn font. Below it, a diagram illustrates the data structure: "variables" (vertical arrows), "observations" (horizontal arrows), and "values" (circles).

#tidyuesday

Tidy Tuesday Screencast: analyzing college major & income data in R

6,263 views

👍 174 🗨️ 1 ➦ SHARE ⋮

Tweeting

A screenshot of a Twitter web interface. The browser address bar shows a URL to a Twitter moment. The navigation bar includes links for Home, Moments (highlighted with a red lightning bolt), Notifications, and Messages, along with a search bar. The main content area displays a tweet from Nic Crane (@nic_crane) dated 22 Nov 2018. The tweet features a code block with R code for using the dplyr group_by_if function. To the left of the code block is a menu with options: Edit, Like, Tweet, and Message. Below the code block is a GIF placeholder. At the bottom of the tweet, engagement metrics show 64 retweets and 241 likes. The tweet text explains the group_by_if function and its use in grouping data by factors.

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



A screenshot of a GitHub Gist page. The browser address bar shows the URL: https://gist.github.com/thisisnic/769baad1aee6ef62495531d291ce6320. The GitHub Gist header is visible with a search bar and navigation links. The user 'thisisnic' is shown with their profile picture and the gist title 'tidyr::unnest.md'. It was created 2 months ago. There are buttons for 'Edit', 'Delete', 'Star', and '1' star. Below the title are tabs for 'Code', 'Revisions 1', and 'Stars 1'. There is an 'Embed' button and a 'Download ZIP' button.

When working with list-columns you can use the parameters to `tidyr::unnest()` to specify whether to keep or drop other list-columns.

A screenshot of the content of the 'tidyr::unnest.md' gist. The title is 'Unnest - unpacking list columns'. The content is R code demonstrating how to use the 'unnest()' function from the 'tidyr' package. The code is as follows:

```
library(dplyr)
library(tidyr)

# Use this to show list columns
glimpse(starwars)

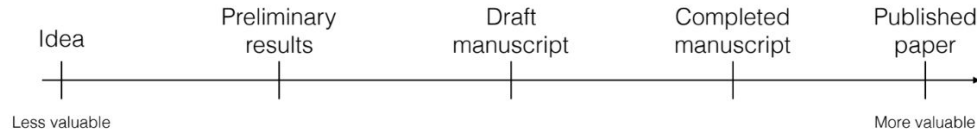
# Unnest 'films' column, drop other list column
by_film <- unnest(starwars, films)
head(by_film)
glimpse(by_film)

# Unnest 'films' column but keep all other list columns
by_film_keep_others <- unnest(starwars, films, .drop = FALSE) # keeps all other list columns
head(by_film_keep_others)
glimpse(by_film_keep_others)
```



“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 