The Lesser Known  s of the Tidyverse

Emily Robinson

@robinson_es

About Me

- Data Analyst at Etsy
- R User for ~6 years
- Enjoy talking about:
  - A/B Testing
  - Building and finding Data Science community
  - R

Disclaimers

This talk represents my own views, not those of Etsy

It’s not Base R vs. Tidyverse

Hadley Wickham @hadleywickham · 30 Aug 2017

Please use as much or as little of the #tidyverse as you feel useful. I want to be as effective in #rstats as possible

15 50 306

Hadley Wickham

@hadleywickham

Following

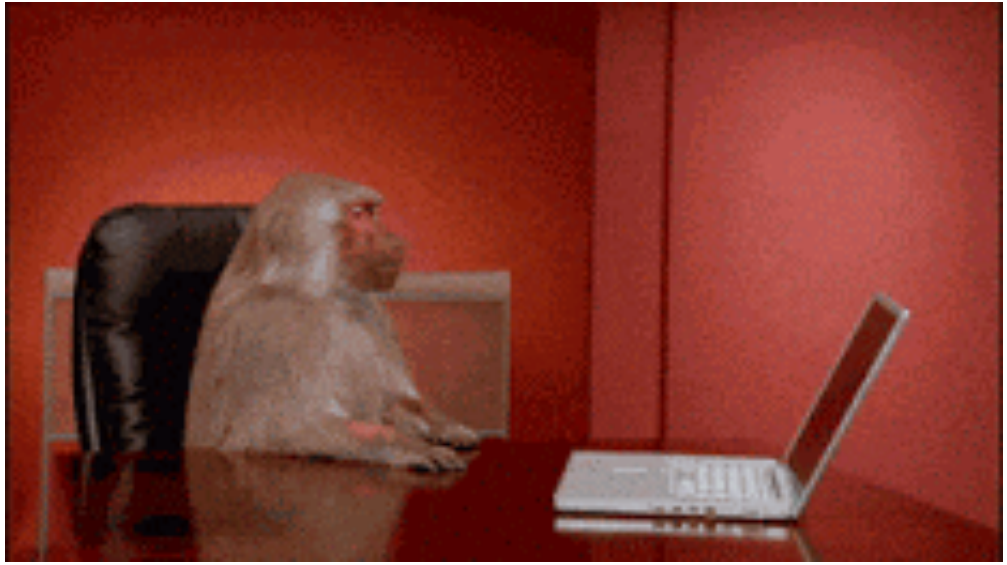
You can not use #tidyverse without base R. It's not a dichotomy. Pick the tools that make you most effective.

8:52 PM - 30 Aug 2017

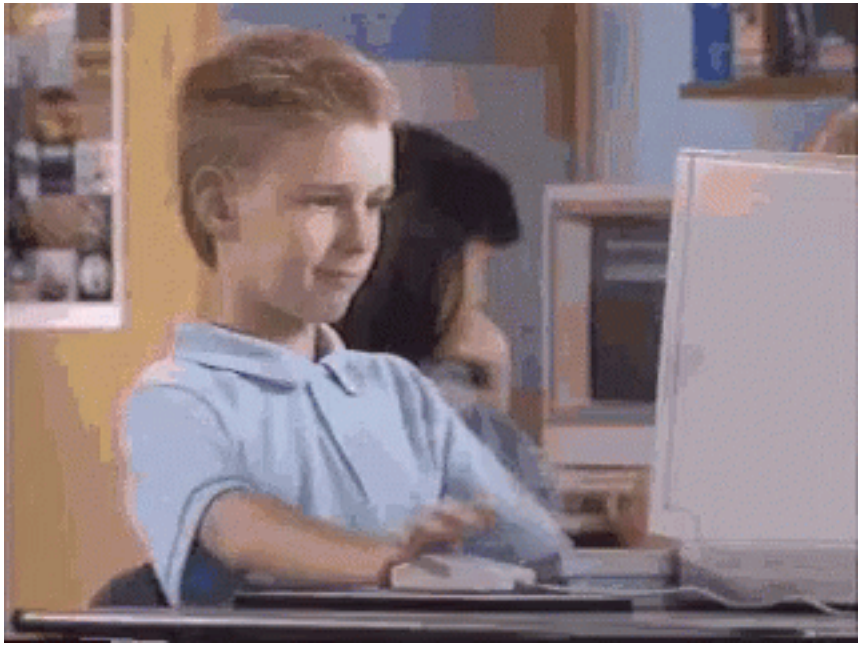
Talk Goals

1. Keep you hip to the lingo

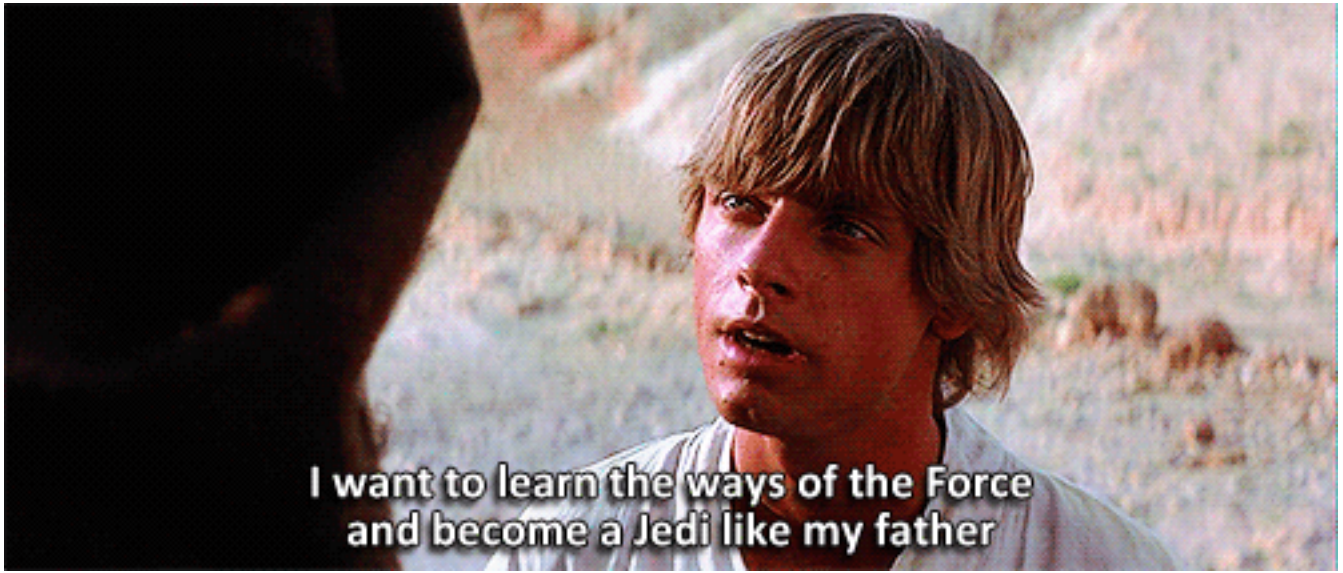
2. Stop you from doing this ...



… by sharing useful functions

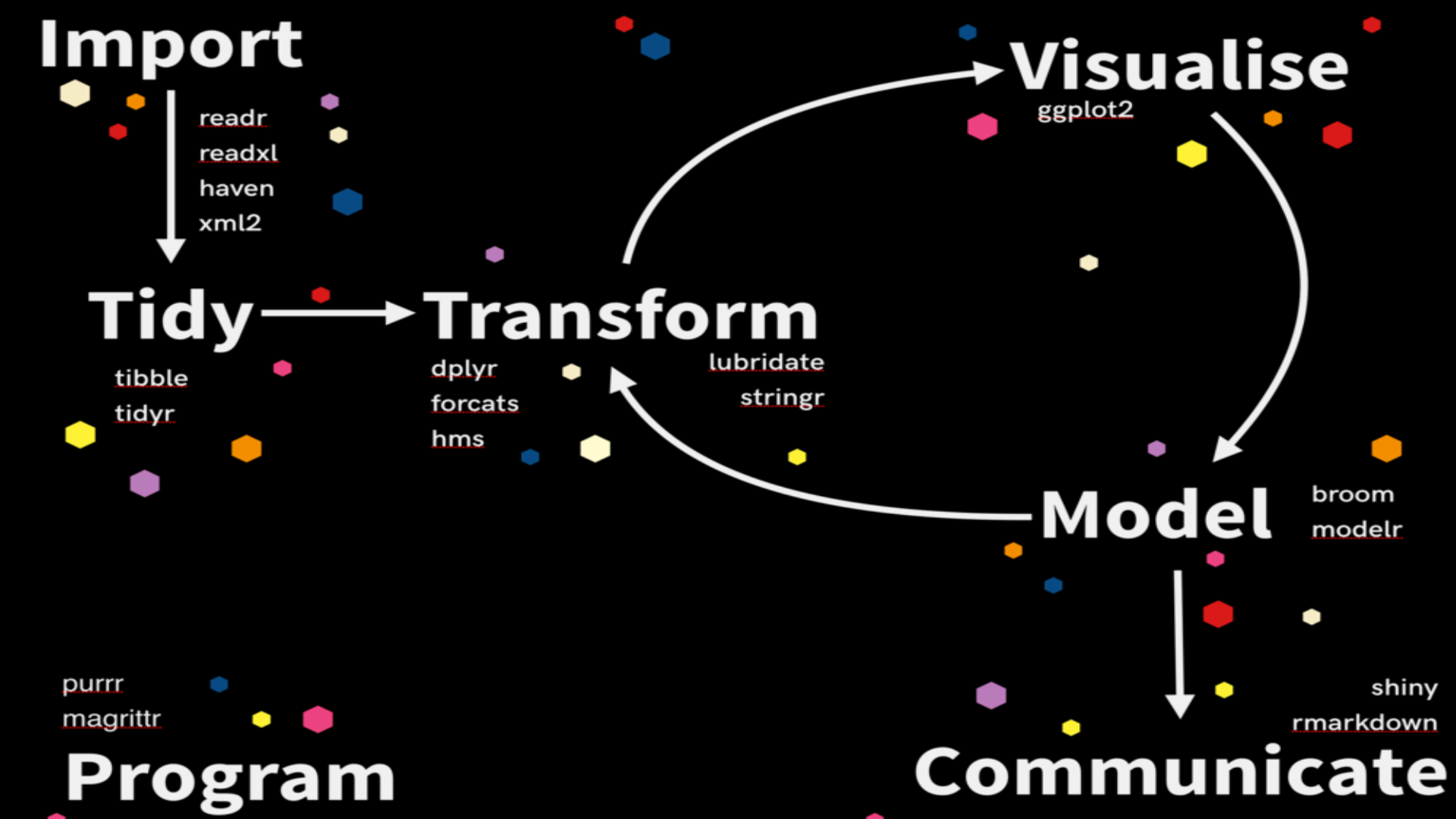


3. Point you to resources



The Tidyverse

An opinionated collection of R packages designed for data science that share an underlying design philosophy, grammar, and data structures


Import
readr
readxl
haven
xml2
Visualise
ggplot2
Tidy
tibble
tidyr
Transform
dplyr
forcats
hms
lubridate
stringr
Model
broom
modelr
purrr
magrittr
Program
shiny
rmarkdown
Communicate

Tidyverse =  ?

Tidyverse =  !

David Robinson
@drob

Following

.@hadleywickham proposes we stop saying "Hadleyverse", start saying "tidyverse" #useR2016 #rstats

9:50 AM - 29 Jun 2016

Tidyverse != Hadleyverse

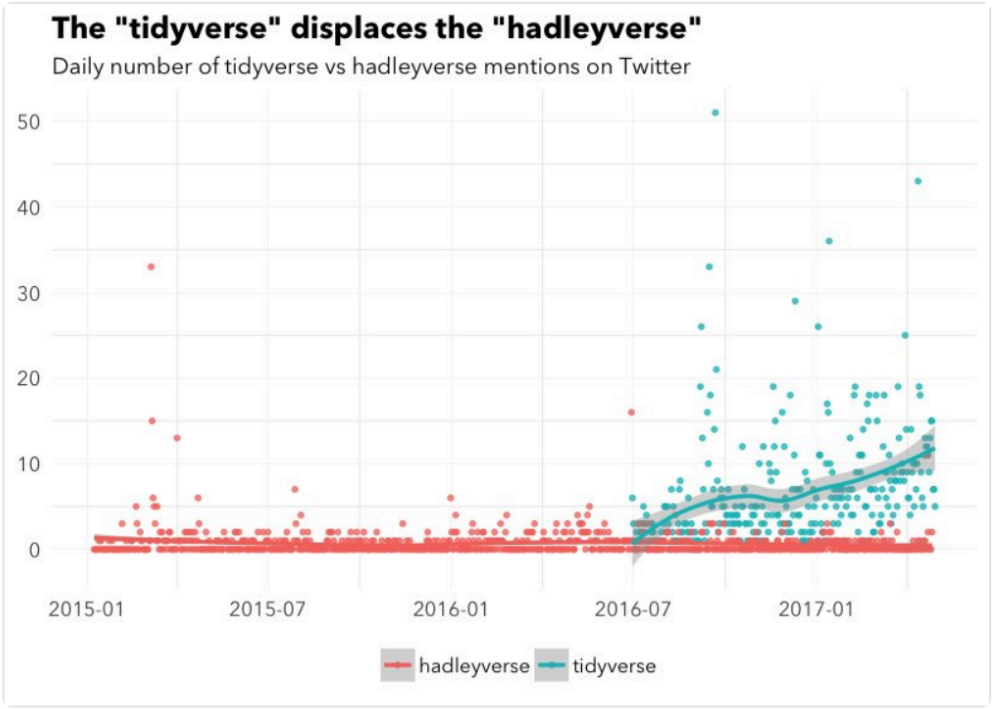
Tidyverse != Hadleyverse

Many other contributors

Mike Kearney 📊
@kearneymw

Following

"Hadleyverse" vs "tidyverse" mentions on Twitter. #rstats



6:35 AM - 19 Sep 2017

Demo

2017

The State of Data Science & Machine Learning

This year, for the first time, we conducted an industry-wide survey to establish a comprehensive view of the state of data science and machine learning. We received over **16,000 responses** and learned a ton about who is working with data, what's happening at the cutting edge of machine learning across industries, and how new data scientists can best break into the field. The below report shares some of our key findings and includes interactive visualizations so you can easily cut the data to find out exactly what you want to know. Here are some sample takeaways:

Step 1: print your dataset!

```
> as.data.frame(multipleChoiceResponses)
                                          GenderSelect                    Country Ag
e                                    EmploymentStatus
1     Non-binary, genderqueer, or gender non-conforming                       <NA>  N
A                                   Employed full-time
2                                             Female               United States  3
0                   Not employed, but looking for work
3                                               Male                     Canada  2
8                   Not employed, but looking for work
4                                               Male               United States  5
6 Independent contractor, freelancer, or self-employed
      StudentStatus                                                 LearningDataScie
nce CodeWriter CareerSwitcher
1              <NA>                                                             <
NA>         Yes            <NA>
2              <NA>                                                             <
NA>        <NA>            <NA>
3              <NA>                                                             <
NA>        <NA>            <NA>
4              <NA>                                                             <
NA>         Yes            <NA>
                     CurrentJobTitleSelect  TitleFit
1                    DBA/Database Engineer      Fine
```

Problem: it takes over the console

Solution: as_tibble()

Prints only 10 rows and the columns that fit on the screen


```
> multipleChoiceResponses
# A tibble: 16,716 x 228
                                       GenderSelect       Country   Age
                                              <chr>         <chr> <int>
 1 Non-binary, genderqueer, or gender non-conforming          <NA>    NA
 2                                            Female United States    30
 3                                              Male        Canada    28
 4                                              Male United States    56
 5                                              Male        Taiwan    38
 6                                              Male        Brazil    46
 7                                              Male United States    35
 8                                            Female         India    22
 9                                            Female     Australia    43
10                                              Male        Russia    33
# ... with 16,706 more rows, and 225 more variables: EmploymentStatus <chr>,
#   StudentStatus <chr>, LearningDataScience <chr>, CodeWriter <chr>,
#   CareerSwitcher <chr>, CurrentJobTitleSelect <chr>, TitleFit <chr>,
#   CurrentEmployerType <chr>, MLToolNextYearSelect <chr>,
#   MLMethodNextYearSelect <chr>, LanguageRecommendationSelect <chr>,
#   PublicDatasetsSelect <chr>, LearningPlatformSelect <chr>,
#   LearningPlatformUsefulnessArxiv <chr>, LearningPlatformUsefulnessBlogs <chr>,
#   LearningPlatformUsefulnessCollege <chr>,
#   LearningPlatformUsefulnessCompany <chr>,
```

Step 2: examine your NAs

```{r}
sum(is.na(multiple_choice_responses_base$StudentStatus))
```

```
[1] 0
```

```{r}
multiple_choice_responses_base %>%
  dplyr::count(StudentStatus)
```

StudentStatus <fctr>	n <int>
	15436
No	299
Yes	981

3 rows

Problem: your NAs aren't actually NAs

Solution: na_if() to replace certain values with NA


```{r}
multiple_choice_responses_base %>%
  na_if("") %>%
  count(StudentStatus)
```

StudentStatus <fctr>	n <int>
No	299
Yes	981
NA	15436

3 rows

Step 3: examine your numeric columns

Problem: how I can I do this quickly?

Solution: dplyr::select_if() + skimr::skim()

+

Skimr

```{r}
multiple_choice_responses %>%
  select_if(is.numeric) %>%
  skimr::skim()
```

```
Skim summary statistics
 n obs: 16716
 n variables: 13

Variable type: integer
          variable missing complete     n  mean    sd p0 p25 median p75 p100     hist
               Age     331    16385 16716 32.37 10.47  0  25     30  37  100 ▃▇▃▁▁▁▁▁
 TimeGatheringData    9186     7530 16716 36.14 21.65  0  20     35  50  100 ▃▇▃▂▁▁▁▁
   TimeOtherSelect    9203     7513 16716  2.4  12.16  0   0      0   0  100 ▇▁▁▁▁▁▁▁

Variable type: numeric
                     variable missing complete     n  mean    sd p0 p25 median p75 p100     hist
       LearningCategoryKaggle    3590    13126 16716  5.53 11.07  0   0      0  10  100 ▇▁▁▁▁▁▁▁
 LearningCategoryOnlineCourses    3590    13126 16716 27.38 26.86  0   5     20  40  100 ▇▃▂▂▁▁▁▁
        LearningCategoryOther    3622    13094 16716  1.8   9.36  0   0      0   0  100 ▇▁▁▁▁▁▁▁
   LearningCategorySelftTaught    3607    13109 16716 33.37 25.79  0  15     30  50  100 ▇▅▃▂▂▁▁▁
```

Step 4: examine a single column

```{r}
multiple_choice_responses %>%
  count(WorkMethodsSelect, sort = TRUE)
```

WorkMethodsSelect <chr>
CNNs
Bayesian Techniques
Data Visualization,Logistic Regression
Data Visualization,Decision Trees
Data Visualization,Logistic Regression,Time Series Analysis
Natural Language Processing
CNNs,Neural Networks
Data Visualization,Text Analytics
A/B Testing,Data Visualization
Logistic Regression,Time Series Analysis

11–20 of 6,191 rows | 1–1 of 2 columns

Problem: it has multiple answers in each row

Solution: stringr::str_split() ...


```{r}
multiple_choice_responses %>%
  mutate(work_method = str_split(WorkMethodsSelect, ",")) %>%
  select(work_method)
```

work_method <list>
<chr [5]>
<chr [12]>
<chr [17]>
<chr [14]>
<chr [12]>
<chr [1]>
<chr [14]>
<chr [12]>
<chr [7]>
<chr [5]>

1–10 of 7,773 rows

Solution: stringr::str_split() and tidyr::unnest()

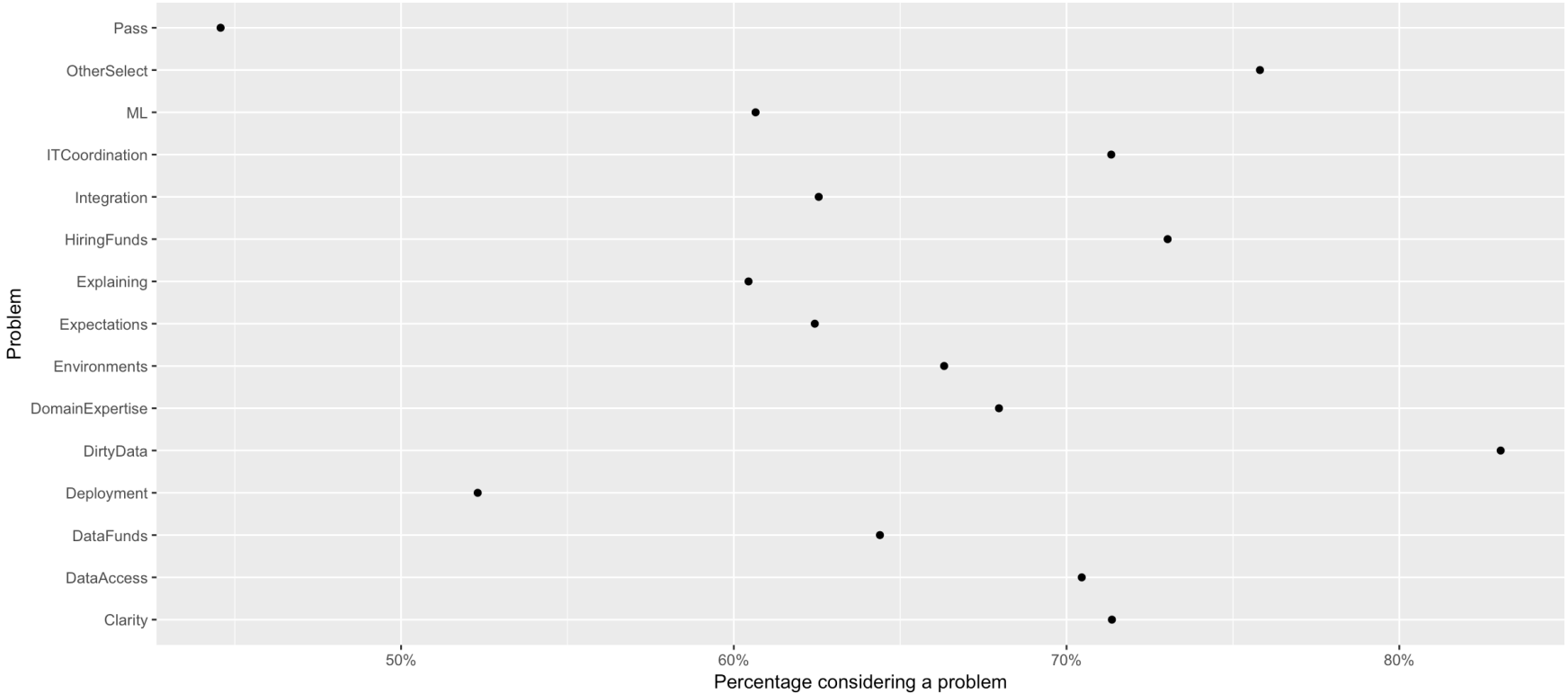
```{r}
multiple_choice_responses %>%
  mutate(work_method = str_split(WorkMethodsSelect, ",")) %>%
  select(work_method) %>%
  unnest()
```

work_method <chr>
Association Rules
Collaborative Filtering
Neural Networks
PCA and Dimensionality Reduction
Random Forests
A/B Testing
Bayesian Techniques
Data Visualization
Decision Trees
Ensemble Methods

1–10 of 59,497 rows

+

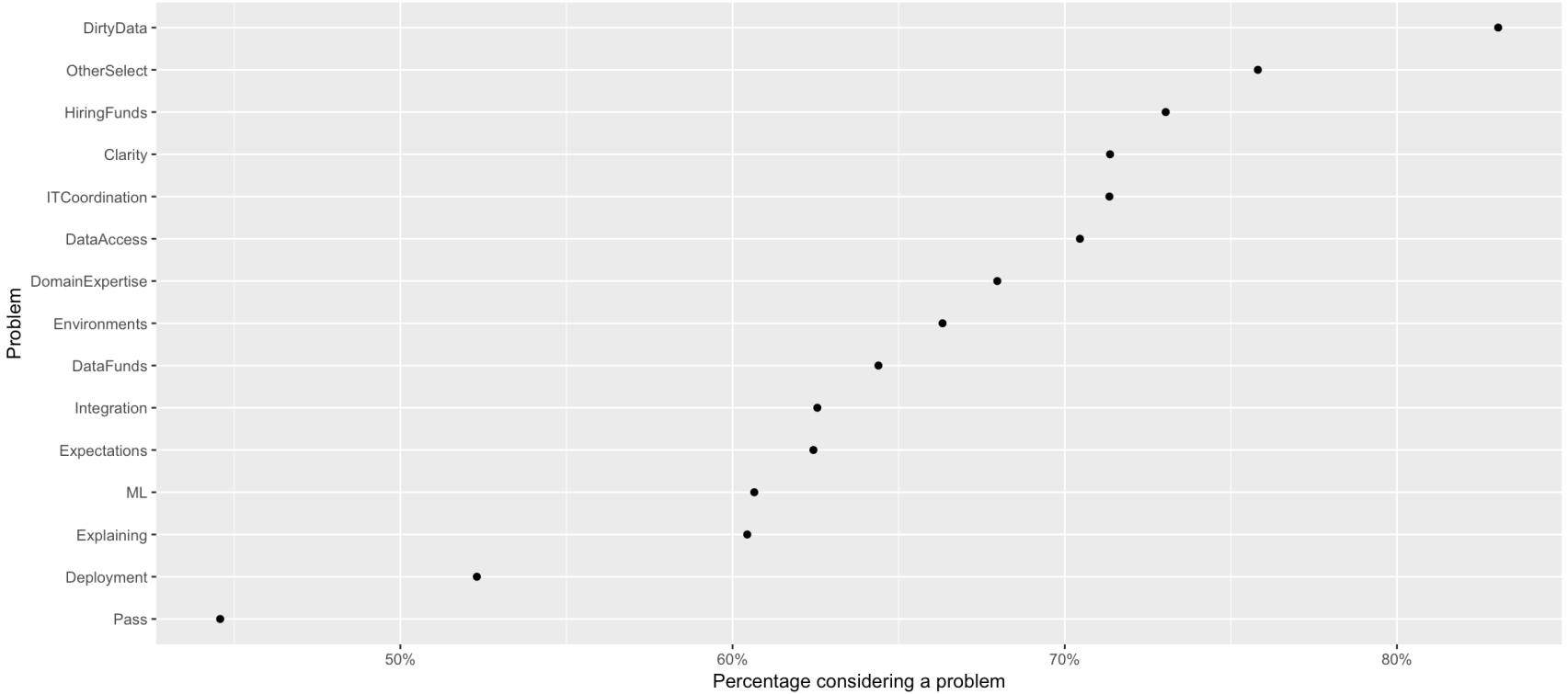
Step 5: make a scatterplot!



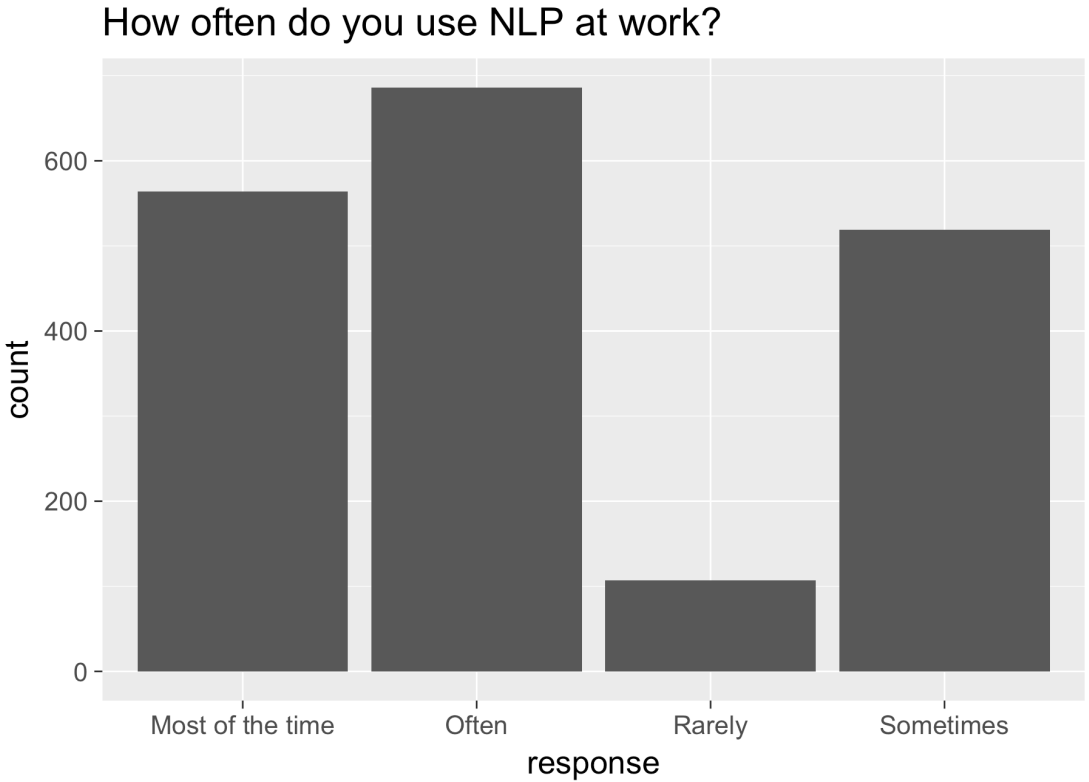
Problem: it’s a mess

Solution: fct_reorder() to order one axis by the other


```
ggplot(WorkChallenges, aes(x = fct_reorder(question,
perc_problem), y = perc_problem)) + geom_point()
```



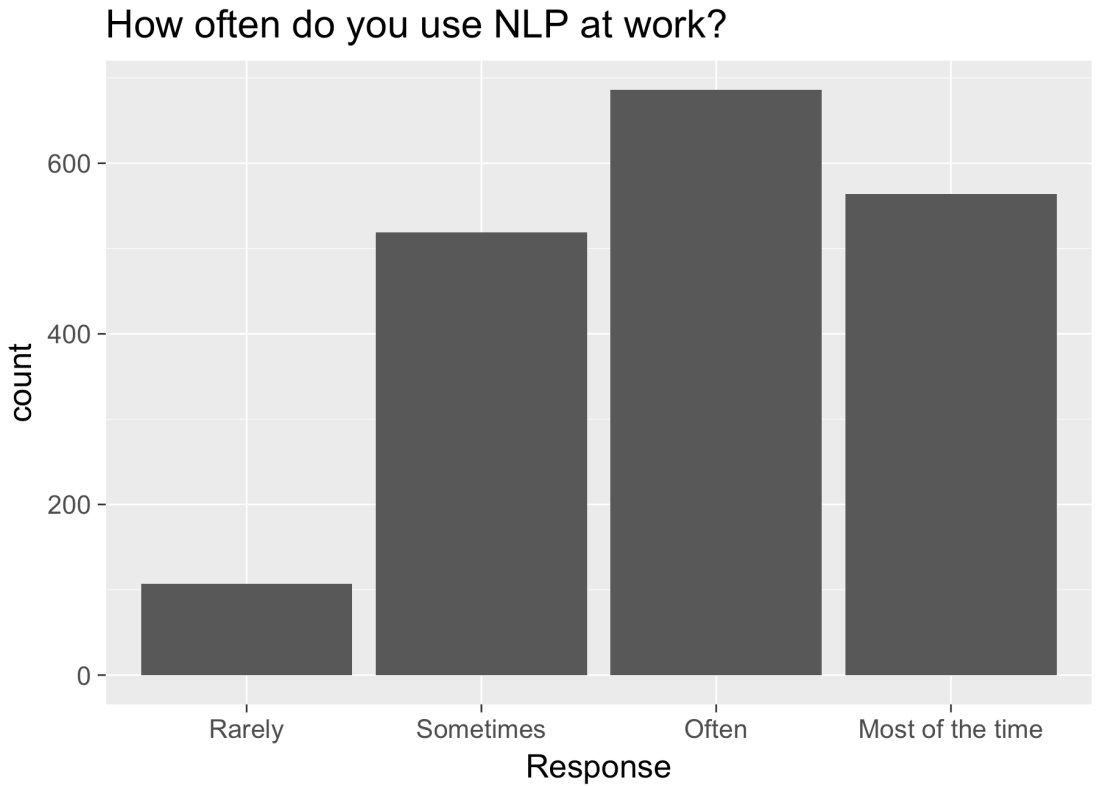
Step 6: make a bar chart!



Problem: your scale is mis-ordered

Solution: fct_relevel() to manually order your factor


```
ggplot(aes(x = fct_relevel(response, "Rarely", "Sometimes",
"Often", "Most of the time"))) + geom_bar()
```



Final step: do something cool and new!

Problem:

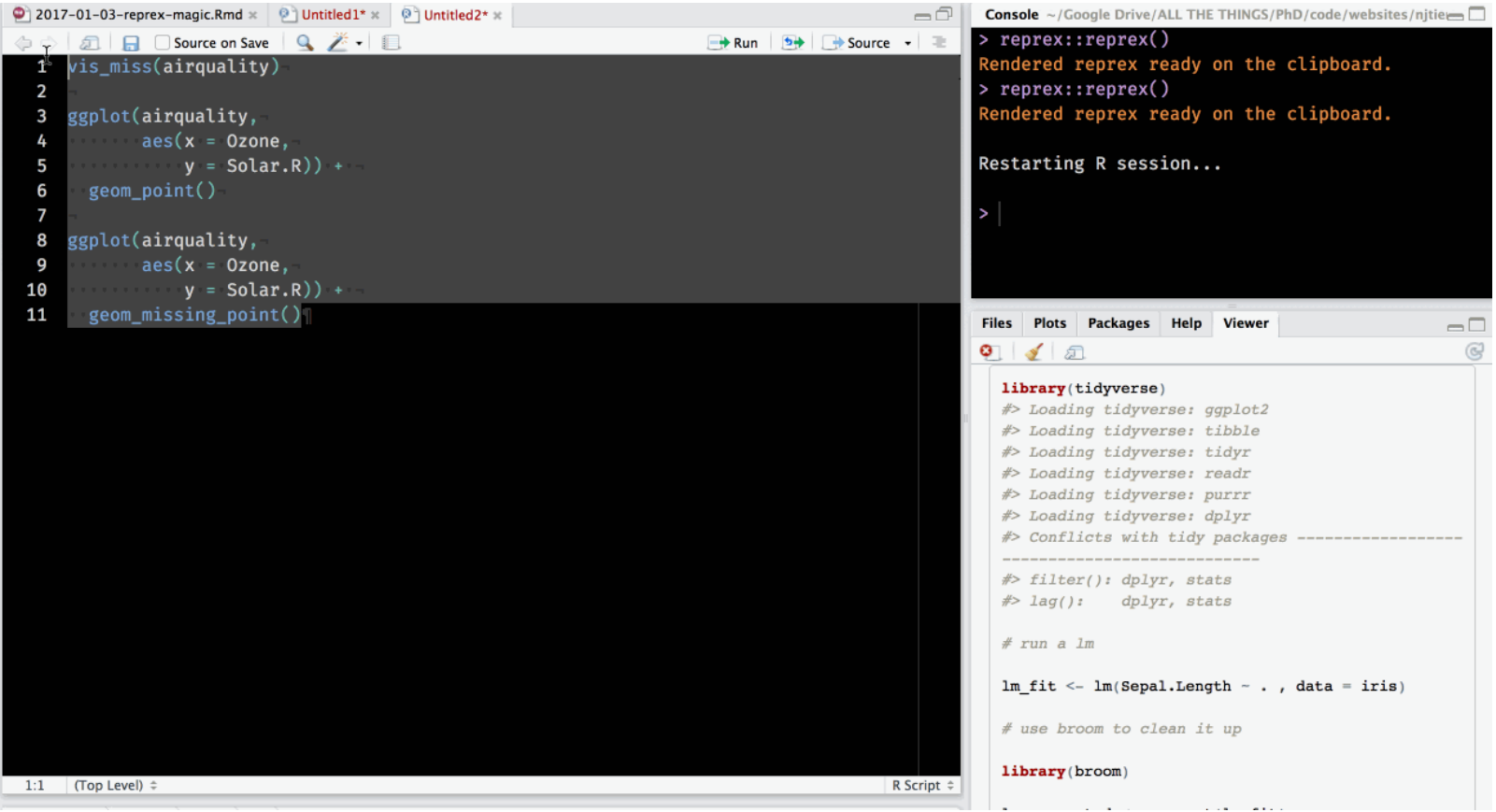
One solution: make a minimal reproducible example

+

Part 0 (optional): use tribble() to make a toy dataset

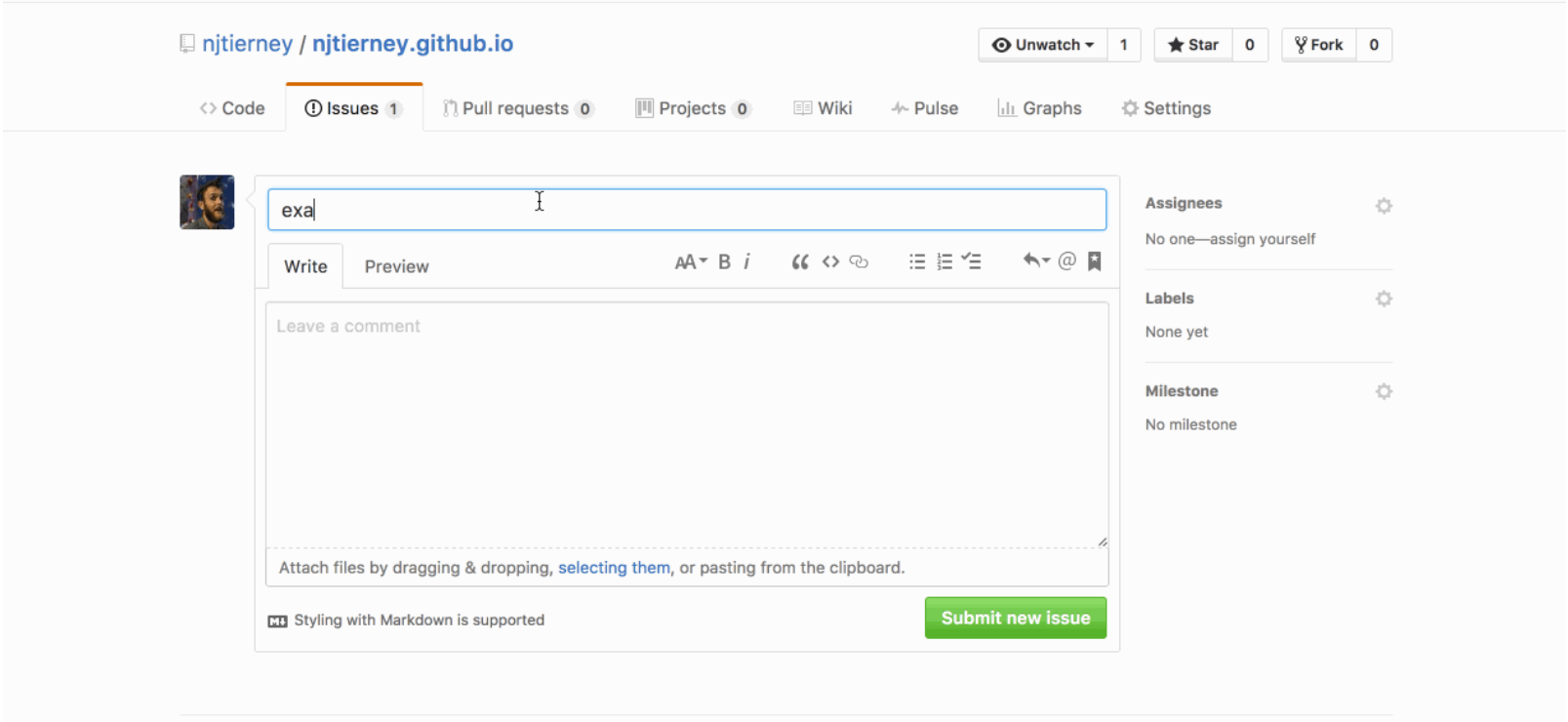

```
library(tibble)
#> Warning: package 'tibble' was built under R version 3.4.1
original_df <- tribble(
  ~date, ~ab_test,  ~ab_variant, ~event_name, ~perc_w_event,
  "10-09-2017", "cool_test",  "old", "clicks", .1,
  "10-09-2017", "cool_test", "new",  "clicks", .2,
  "10-09-2017", "cool_test", "third_variant",  "clicks", .5,
  "10-09-2017", "awesome_test", "off", "clicks", .3,
  "10-09-2017", "awesome_test", "on",  "clicks", .4,
  "10-10-2017", "awesome_test", "off",  "clicks", .6,
  "10-10-2017", "awesome_test", "on",  "clicks", .8

)
```

Part 1: Use reprex() to find any problems



Credit: Nick Tiernay, https://www.njtierney.com/post/2017/01/11/magic-reprex/

Part 2: Use reprex() to post your question or issue



Credit: Nick Tiernay, https://www.njtierney.com/post/2017/01/11/magic-reprex/

Review

tibble::as_tibble

tibble::tribble

dplyr::na_if

dplyr::select_if

skimr::skim

stringr::str_split

tidyr::unnest

forcats::fct_reorder

forcats::fct_relevel

reprex::reprex

Resources

R for Data Science

Garrett Grolemund

Hadley Wickham

Welcome

This is the website for **“R for Data Science”**. This book will teach you how to do data science with R: You’ll learn how to get your data into R, get it into the most useful structure, transform it, visualise it and model it. In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you’ll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You’ll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. You’ll also learn how to manage cognitive resources to facilitate discoveries when wrangling, visualising, and exploring data.

#rstats Twitter

Emily Robinson
@robinson_es

#rstats twitter: is there a way to globally set the color scale for ggplot2? theme_set() seems to work only for background.

7:15 AM - 20 Dec 2017

#rstats Twitter

Ildi Czeller
@czeildi

Following

Replying to @robinson_es

Not native ggplot2, but I use {ggthemr}, it provides define_palette, ggthemr(your_palette) to set palette for all plots and ggthemr_reset to go back to defaults. Concrete example: ildiczeller.com/2017/10/15/cus …

Datacamp.com

tidyverse

All Technologies All Topics

All Data Science Courses

Introduction to the Tidyverse

Get started on the path to exploring and visualizing your own data with the tidyverse, a powerful and popular collect...

Continue Course

Communicating with Data in the Tidyverse

Leverage the power of tidyverse tools to create publication-quality graphics and custom-styled reports that communica...

4 hours

TIMO GROSSENBACHER
Data Journalist at SRF Data

Working with Dates and Times in R

Learn the essentials of parsing, manipulating and computing with dates and times in R.

4 hours **Play preview**

CHARLOTTE WICKHAM
Assistant Professor at Oregon State University

Base R to Tidyverse Translation

www.significantdigits.org/2017/10/switching-from-base-r-to-tidyverse/

Base R command	Tidyverse Command	What it does and why you should use the tidyverse version	Comment
read.csv()	read_csv()	reads in a csv file, but its much faster, shows progress bar for large files, can automatically parse data types	also see read_delim(), read_tsv() and readxl::read_xlsx()

And much more!

- Tidyverse.org
- community.rstudio.com/c/tidyverse
- https://www.rstudio.com/resources/cheatsheets/
- https://medium.com/@kierisi/r4ds-the-next-iteration-d51e0a1b0b82

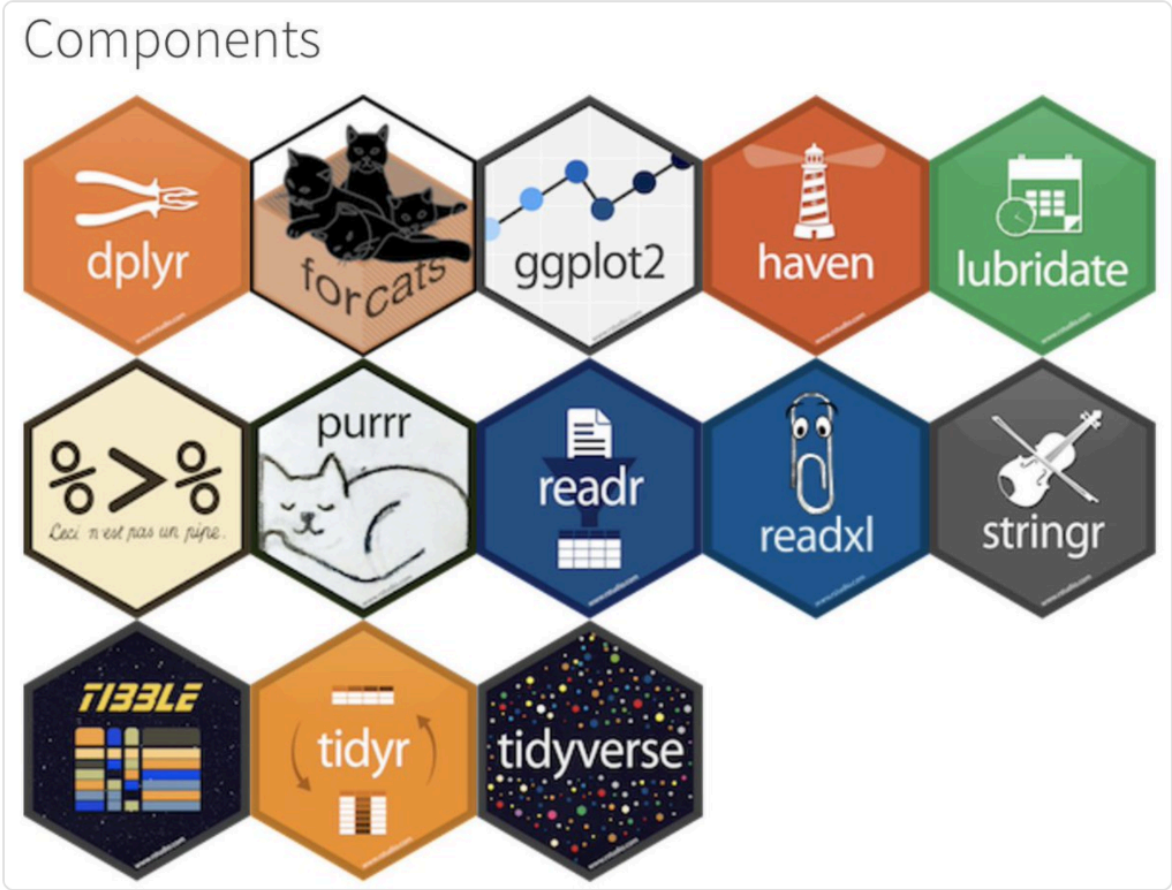
Christine Zhang @christinezhang · Jun 5

TIL that there are packages called "forcats" & "purrr" in @hadleywickham's **tidyverse** #rstats #**hadleyverse** tidyverse.org

The tidyverse

Come for the stickers and package names ...

Stay for the friendly community and happy workflow.



Thank You!

tiny.cc/rstudiotalk

robinsones.github.io

@robinson_es