Tidyverse

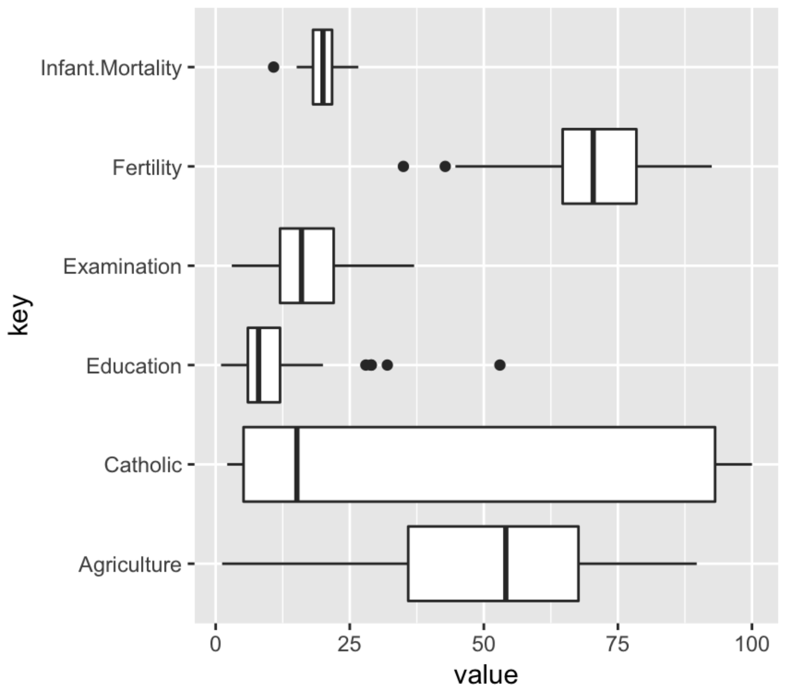
What is the “tidyverse”?

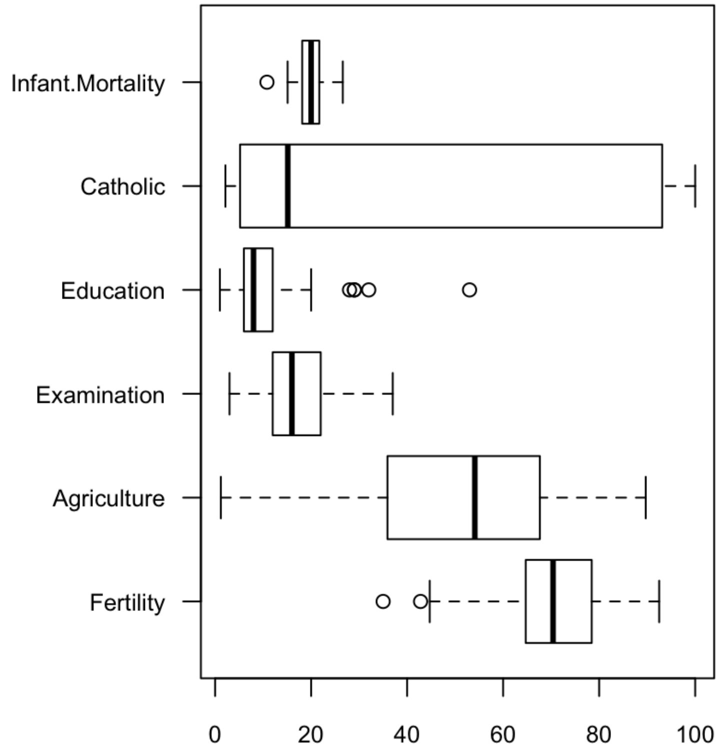
- “an opinionated collection of R packages designed for data science. All packages share an underlying philosophy and common APIs.”
- formerly referred to as the “hadleyverse” for Hadley Wickham
- packages are strongly associated with RStudio, but not exclusively

Core tidyverse packages

- ggplot2
- dplyr
- tidyr
- readr
- purrr
- tibble
- stringr
- forcats

10 differences between the tidyverse and base R





1. Base R

Base:

```r
barplot(1:5, horiz = TRUE)
boxplot(1:5, horizontal = TRUE)
```

1. Tidyverse is … more consistent

Tidyverse:

```
ggplot(...) + geom_bar() + coord_flip()
ggplot(...) + geom_boxplot() + coord_flip()
```

2. Base R

Base:

```r
df <- data.frame(x = 1:4, y = 1:2)
```

2. Base R

Base:

```r
df <- data.frame(x = 1:4, y = 1:2)
```

```r
df
```

```
##   x y
## 1 1 1
## 2 2 2
## 3 3 1
## 4 4 2
```

2. Tidyverse … fails faster

Tidyverse:

```
df <- tibble(x = 1:4, y = 1:2)
```

2. Tidyverse … fails faster

Tidyverse:

```
library(tibble)
df <- tibble(x = 1:4, y = 1:2)
```

```
## Tibble columns must have consistent lengths, only values of length one are recycled:
## * Length 2: Column `y`
## * Length 4: Column `x`
```

3. Base R

Base:

```r
df <- read.csv("animals.csv")
df
```

```
##     animal count
## 1 elephant     3
## 2      cat     2
## 3     frog     6
```

```r
str(df)
```

```
## 'data.frame':    3 obs. of  2 variables:
##  $ animal: Factor w/ 3 levels "cat","elephant",..: 2 1 3
##  $ count : int  3 2 6
```

3. Tidyverse … avoids factors

Tidyverse:

```r
library(readr)
df <- read_csv("animals.csv")
str(df)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 3 obs. of  2 variables:
##  $ animal: chr  "elephant" "cat" "frog"
##  $ count : num  3 2 6
##  - attr(*, "spec")=
##   .. cols(
##   ..   animal = col_character(),
##   ..   count = col_double()
##   .. )
```

4. Base R

Base:

```r
df <- read.csv("animals.csv")
df
```

```
##     animal count
## 1 elephant     3
## 2      cat     2
## 3     frog     6
```

```r
x <- cbind(df[,1], df[,2])
class(x)
```

4. Base R

Base:

```
df <- read.csv("animals.csv")
df
```

```
##     animal count
## 1 elephant     3
## 2      cat     2
## 3     frog     6
```

```
x <- cbind(df[,1], df[,2])
class(x)
```

```
## [1] "matrix"
```

```
x
```

4. Base R

Base:

```
df <- read.csv("animals.csv")
df
```

```
##      animal count
## 1 elephant     3
## 2      cat     2
## 3     frog     6
```

```
x <- cbind(df[,1], df[,2])
class(x)
```

```
## [1] "matrix"
```

```
x
```

```
##      [,1] [,2]
## [1,]    2    3
## [2,]    1    2
## [3,]    3    6
```

4. Tidyverse … is more predictable

Tidyverse:

```
tib <- read_csv("animals.csv")
tib
```

```
## # A tibble: 3 x 2
##   animal   count
##   <chr>    <dbl>
## 1 elephant     3
## 2 cat          2
## 3 frog         6
```

```
x <- bind_cols(tib[,1], tib[,2])
class(x)
```

4. Tidyverse … is more predictable

Tidyverse:

```
tib <- read_csv("animals.csv")
tib
```

```
## # A tibble: 3 x 2
##   animal   count
##   <chr>    <dbl>
## 1 elephant     3
## 2 cat          2
## 3 frog         6
```

```
x <- bind_cols(tib[,1], tib[,2])
class(x)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
x
```

4. Tidyverse … is more predictable

Tidyverse:

```
tib <- read_csv("animals.csv")
tib
```

```
## # A tibble: 3 x 2
##   animal   count
##   <chr>    <dbl>
## 1 elephant     3
## 2 cat          2
## 3 frog         6
```

```
x <- bind_cols(tib[,1], tib[,2])
class(x)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
x
```

```
## # A tibble: 3 x 2
##   animal   count
##   <chr>    <dbl>
## 1 elephant     3
## 2 cat          2
## 3 frog         6
```

4. Base R

Base:

```r
class(df)
```

```
## [1] "data.frame"
```

```r
class(df[,1])
```

```
## [1] "factor"
```

4. Base R

Base:

```
class(df)
```

```
## [1] "data.frame"
```

```
class(df[,1])
```

```
## [1] "factor"
```

4. Tidyverse is … more predictable

Tidyverse:

```r
class(tib)
```

```
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

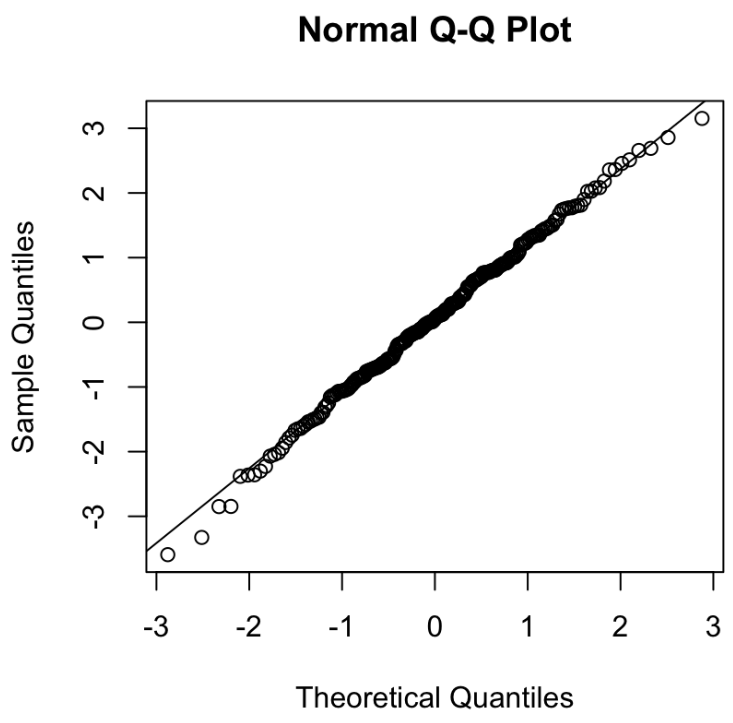
```r
class(tib[,1])
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

5. Base R

Base:

```
# p. 115, Modern Applied Statistics with S-Plus (1999)
x <- rt(250, 9)
qqnorm(x); qqline(x)
```



Source: Venables and Ripley, *Modern Applied Statistics with S-Plus* (1999), p. 115.

5. Tidyverse is …

Tidyverse:

```r
df <- iris %>% dplyr::add_rownames()
```

5. Tidyverse is … still evolving

Tidyverse:

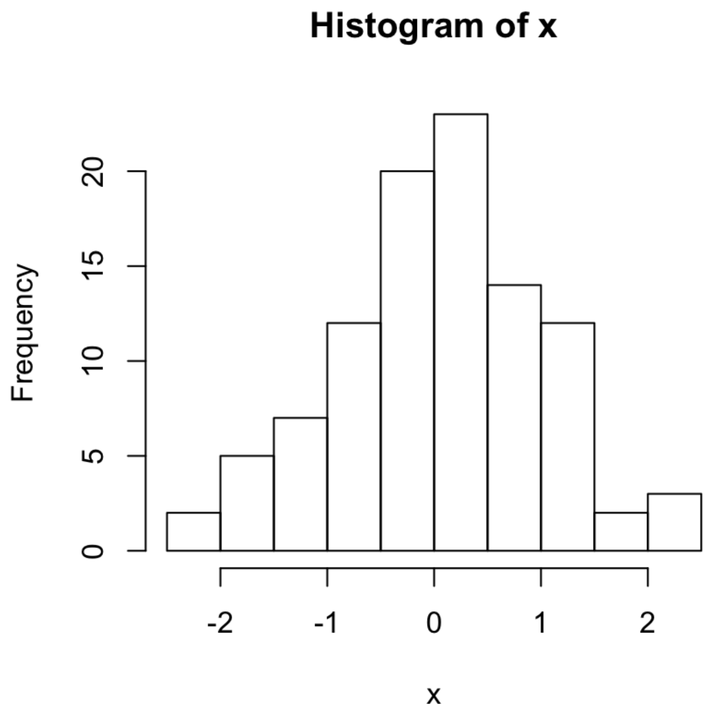
```
df <- iris %>% dplyr::add_rownames()
```

```
## Warning: Deprecated, use tibble::rownames_to_column() instead.
```

6. Base R

Base:

```r
x <- rnorm(100)
hist(x)
```



6. Tidyverse … avoids vectors

Tidyverse:

```
library(ggplot2)
ggplot(x, aes(x)) + geom_histogram()
```

```
## Error: `data` must be a data frame, or other object coercible by `fortify()`, not a numeric vector
```

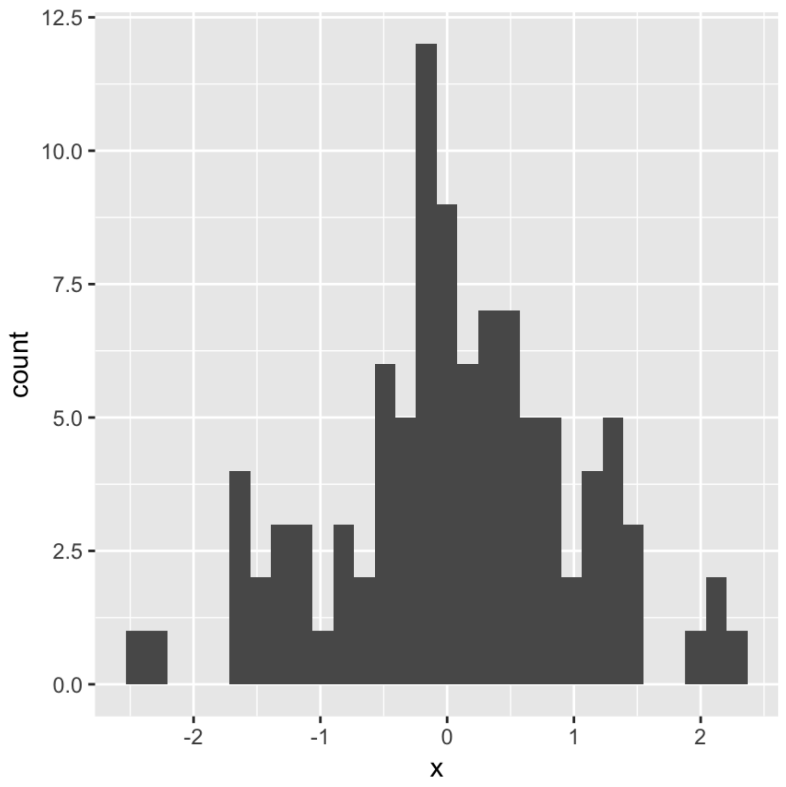
6. Tidyverse … avoids vectors

Tidyverse:

```r
library(ggplot2)
ggplot(x, aes(x)) + geom_histogram()
```

```
## Error: `data` must be a data frame, or other object coercible by `fortify()`, not a numeric vector
```

```r
ggplot(data.frame(x), aes(x)) + geom_histogram()
```

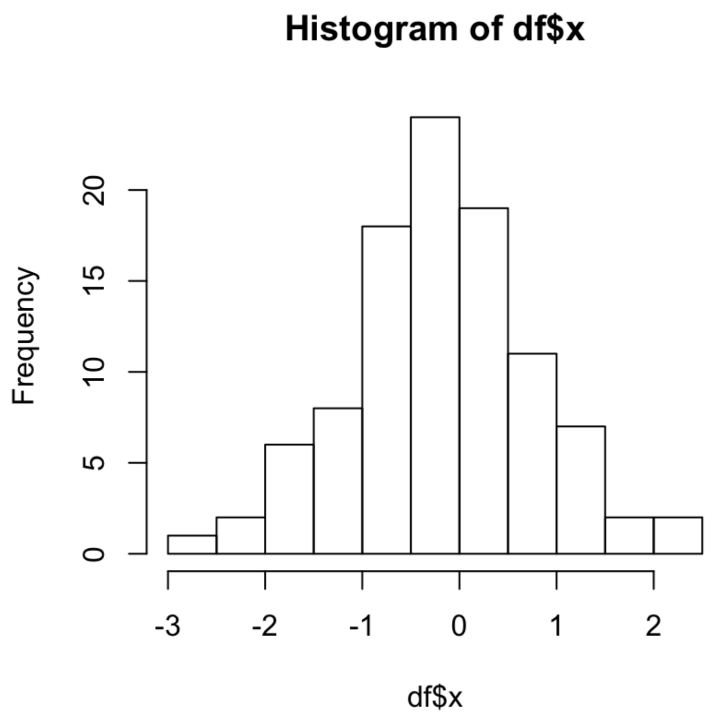


12.5
10.0
7.5
5.0
2.5
0.0
count
-2
-1
0
1
2
x

7. Base R

Base:

```r
df <- data.frame(x = rnorm(100))
hist(df$x)
```

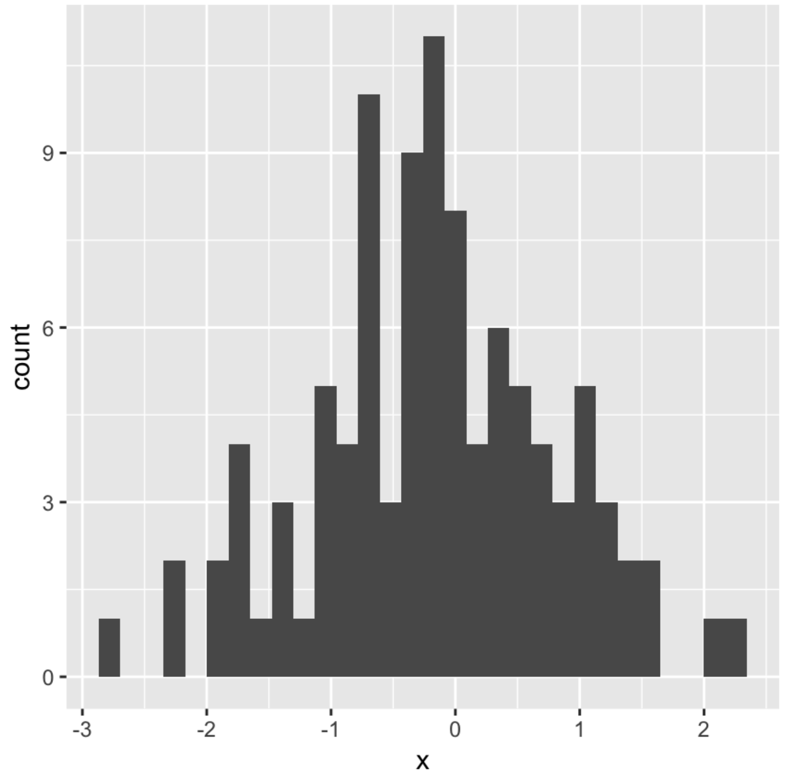


7. Tidyverse is … more talkative

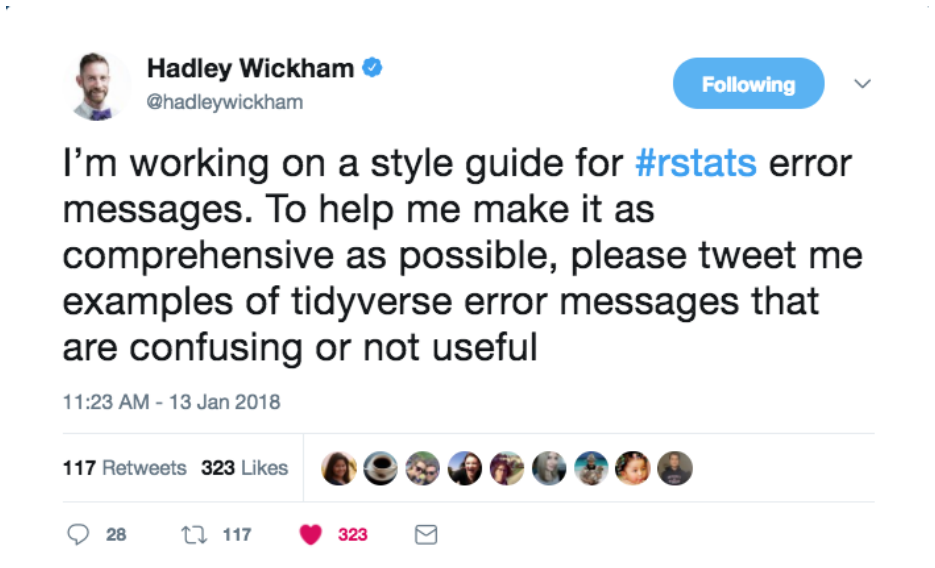
Tidyverse:

```
library(ggplot2)
ggplot(df, aes(x)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



7. Tidyverse is … more talkative


Hadley Wickham
@hadleywickham

Following

I'm working on a style guide for #rstats error messages. To help me make it as comprehensive as possible, please tweet me examples of tidyverse error messages that are confusing or not useful

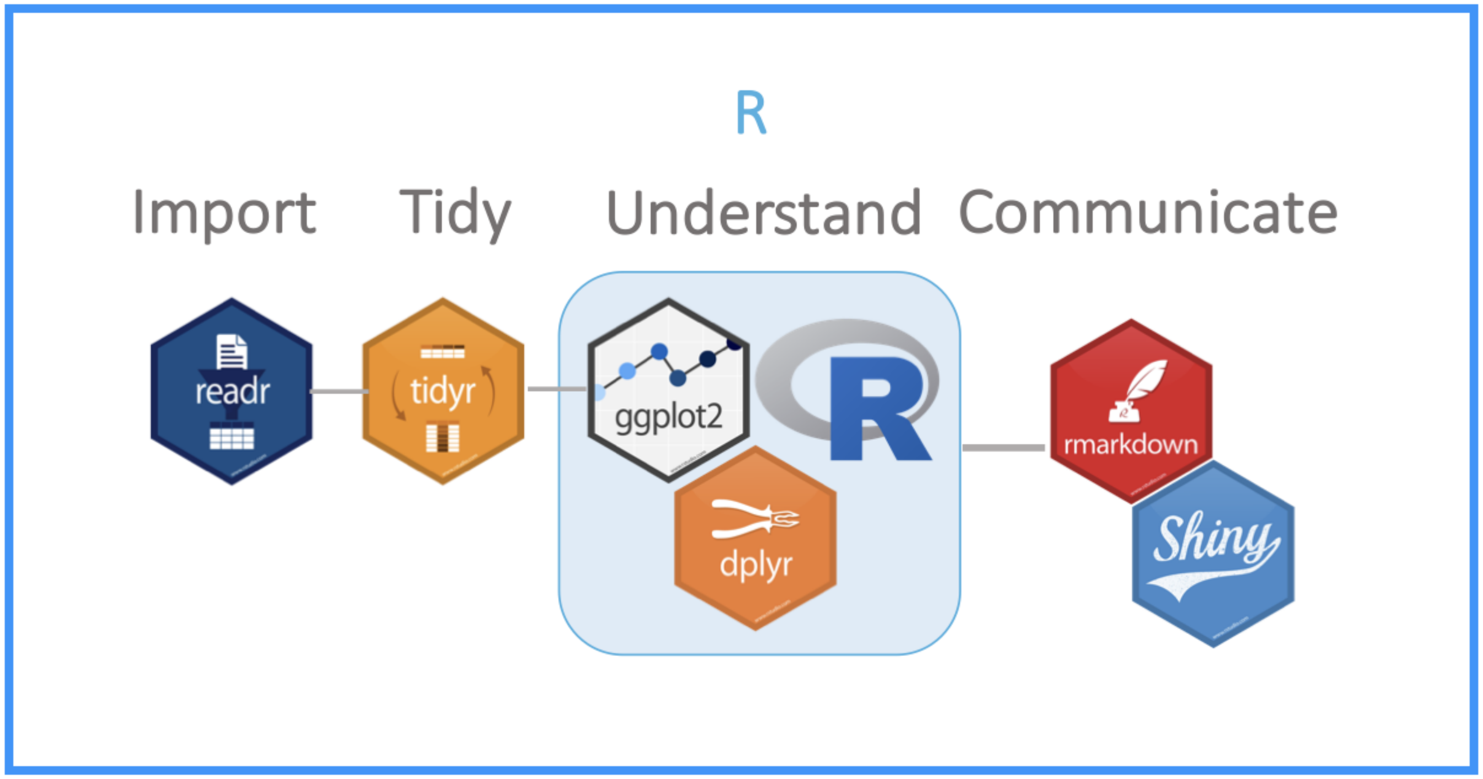
11:23 AM - 13 Jan 2018

117 Retweets 323 Likes

28 117 323

https://twitter.com/hadleywickham/status/952259891342794752

8. Tidyverse is .. more coordinated across tasks



Source: RStudio, https://github.com/rstudio/meetup_roadshow/blob/master/2017%20Meetup%20Roadshow.pptx

9. Tidyverse is … easier for beginners

```r
# base R
crime.by.state <- read.csv("CrimeStatebyState.csv")
crime.ny.2005 <- crime.by.state[crime.by.state$Year==2005 &
                                  crime.by.state$State=="New
  York",
                                c("Type.of.Crime", "Count")]
crime.ny.2005 <- crime.ny.2005[order(crime.ny.2005$Count,
                                     decreasing=TRUE), ]
crime.ny.2005$Proportion <- crime.ny.2005$Count /
                            sum(crime.ny.2005$Count)
summary1 <- aggregate(Count ~ Type.of.Crime,
                      data=crime.ny.2005, FUN=sum)
summary2 <- aggregate(Count ~ Type.of.Crime,
                      data=crime.ny.2005, FUN=length)
final <- merge(summary1, summary2, by="Type.of.Crime")
```

```r
# dplyr
crime.by.state <- read.csv("CrimeStatebyState.csv")
final <- crime.by.state %>%
           filter(State=="New York", Year==2005) %>%
           arrange(desc(Count)) %>%
           select(Type.of.Crime, Count) %>%
           mutate(Proportion=Count/sum(Count)) %>%
           group_by(Type.of.Crime) %>%
           summarise(num.types = n(), counts =
  sum(Count))
```

Goodbye $ [ ]

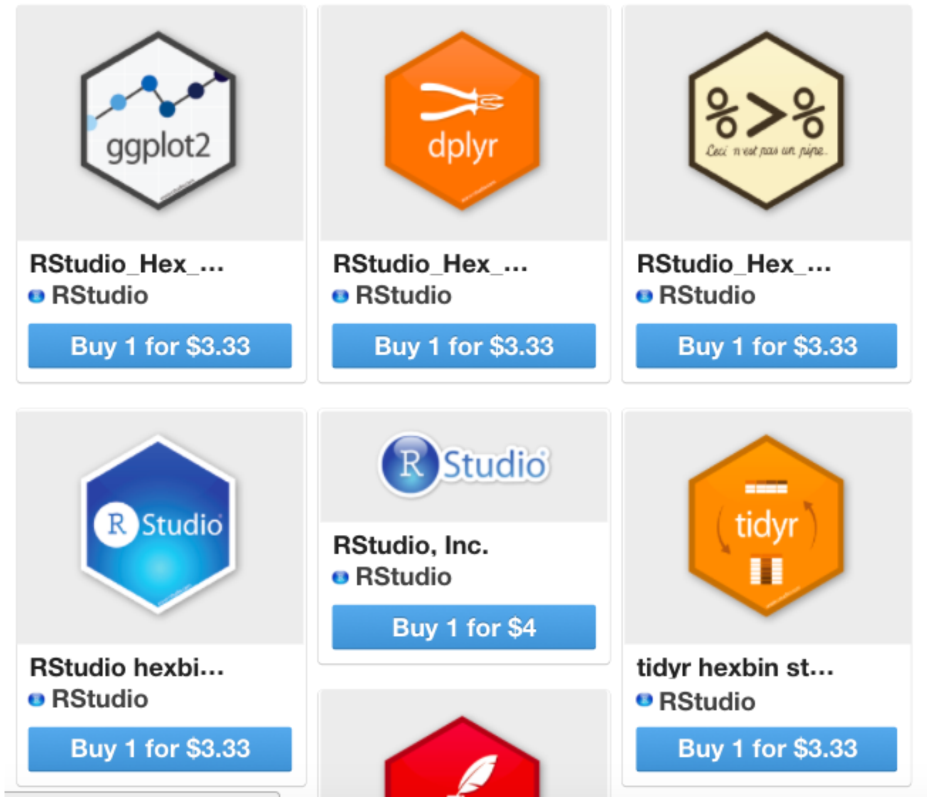
Source: "How `dplyr` replaced my most common R idioms"

http://www.onthelambda.com/2014/02/10/how-dplyr-replaced-my-most-common-r-idioms/

(highly recommended!)

10. Tidyverse … is more collaborative

https://twitter.com/jtrnyc/status/954148122724392960



RStudio_Hex_…
RStudio
Buy 1 for $3.33

RStudio_Hex_…
RStudio
Buy 1 for $3.33

RStudio_Hex_…
RStudio
Buy 1 for $3.33

RStudio hexbi…
RStudio
Buy 1 for $3.33

RStudio, Inc.
RStudio
Buy 1 for $4

tidyr hexbin st…
RStudio
Buy 1 for $3.33

Source: https://www.stickermule.com/user/1070448958/stickers