

Tidy evaluation:

Programming with ggplot2 and dplyr

January 2019

Hadley Wickham

@hadleywickham

Chief Scientist, RStudio



Writing functions

Rule of three: make a function if you've copy-pasted threes times

```
(df$a - min(df$a)) / (max(df$a) - min(df$a))
```

```
(df$b - min(df$b)) / (max(df$b) - min(df$b))
```

```
(df$c - min(df$c)) / (max(df$c) - min(df$c))
```

```
(df$d - min(df$d)) / (max(df$d) - min(df$c))
```

Rule of three: make a function if you've copy-pasted threes times

```
(df$a - min(df$a)) / (max(df$a) - min(df$a))
```

```
(df$b - min(df$b)) / (max(df$b) - min(df$b))
```

```
(df$c - min(df$c)) / (max(df$c) - min(df$c))
```

```
(df$d - min(df$d)) / (max(df$d) - min(df$c))
```

Rule of three: make a function if you've copy-pasted threes times

```
(df$a - min(df$a)) / (max(df$a) - min(df$a))
```

```
(df$b - min(df$b)) / (max(df$b) - min(df$b))
```

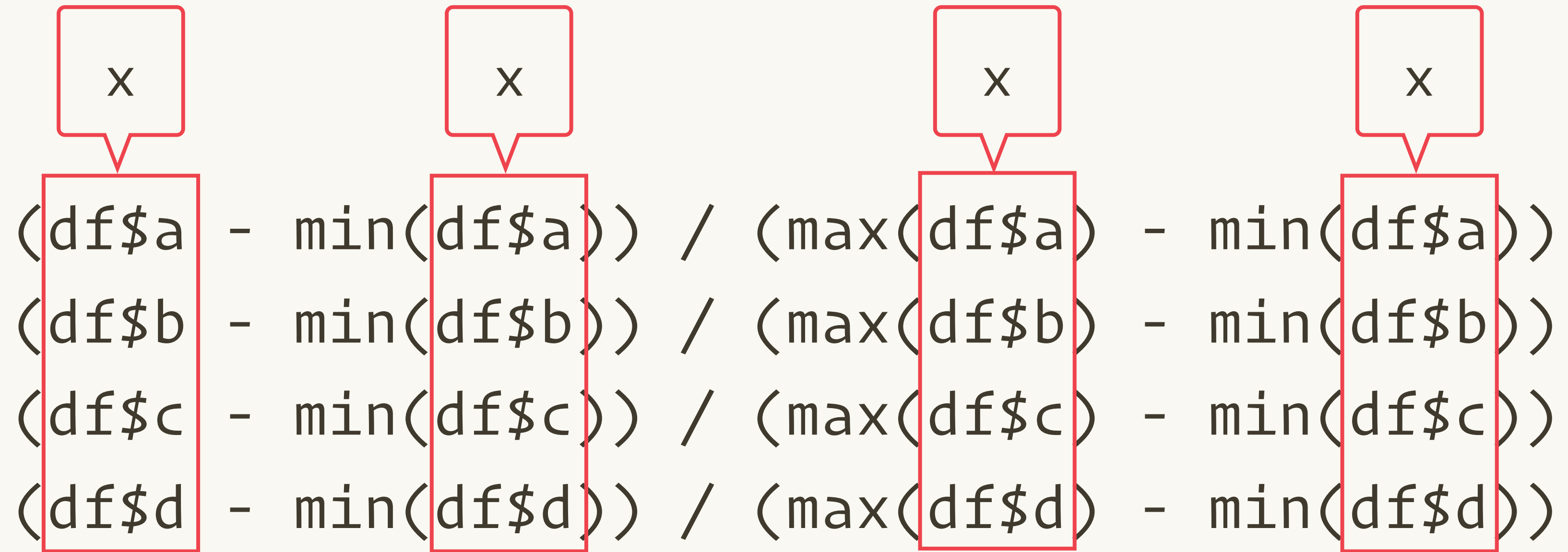
```
(df$c - min(df$c)) / (max(df$c) - min(df$c))
```

```
(df$d - min(df$d)) / (max(df$d) - min(df$d))
```

First, identify the parts that might change

$(df\$a - \min(df\$a)) / (\max(df\$a) - \min(df\$a))$
 $(df\$b - \min(df\$b)) / (\max(df\$b) - \min(df\$b))$
 $(df\$c - \min(df\$c)) / (\max(df\$c) - \min(df\$c))$
 $(df\$d - \min(df\$d)) / (\max(df\$d) - \min(df\$d))$

Then give them names



The diagram shows four identical mathematical expressions arranged horizontally. Each expression is enclosed in a red rectangular box. Above each box is a red speech bubble containing the letter 'x'. The expressions are as follows:

$$\begin{aligned} & \left(\text{df\$a} - \min(\text{df\$a}) \right) / \left(\max(\text{df\$a}) - \min(\text{df\$a}) \right) \\ & \left(\text{df\$b} - \min(\text{df\$b}) \right) / \left(\max(\text{df\$b}) - \min(\text{df\$b}) \right) \\ & \left(\text{df\$c} - \min(\text{df\$c}) \right) / \left(\max(\text{df\$c}) - \min(\text{df\$c}) \right) \\ & \left(\text{df\$d} - \min(\text{df\$d}) \right) / \left(\max(\text{df\$d}) - \min(\text{df\$d}) \right) \end{aligned}$$

Make the function template

```
rescale01 <- function(x) {  
  
}
```


Then copy in one example

```
rescale01 <- function(x) {  
  (df$a - min(df$a)) / (max(df$a) - min(df$a))  
}
```

And use the variable

```
rescale01 <- function(x) {  
  (x - min(x)) / (max(x) - min(x))  
}
```

And maybe refactor a little

```
rescale01 <- function(x) {  
  rng <- range(x)  
  (x - rng[1]) / (rng[2] - rng[1])  
}
```

And handle more cases

```
rescale01 <- function(x) {  
  rng <- range(x, na.rm = TRUE, finite = TRUE)  
  (x - rng[1]) / (rng[2] - rng[1])  
}
```

Rule of three: make a function if you've copy-pasted threes times

```
(df$a - min(df$a)) / (max(df$a) - min(df$a))
```

```
(df$b - min(df$b)) / (max(df$b) - min(df$b))
```

```
(df$c - min(df$c)) / (max(df$c) - min(df$c))
```

```
(df$d - min(df$d)) / (max(df$d) - min(df$d))
```

Rule of three: make a function if you've copy-pasted threes times

```
rescale01(df$a)
```

```
rescale01(df$b)
```

```
rescale01(df$c)
```

```
rescale01(df$d)
```

Why create a function? Because a function:

1. Prevents inconsistencies
2. Emphasises what varies
3. Makes change easier
4. Can have informative name

Motivation

Let's try with some dplyr code

```
df %>% group_by(x1) %>% summarise(mean = mean(y1))
```

```
df %>% group_by(x2) %>% summarise(mean = mean(y2))
```

```
df %>% group_by(x3) %>% summarise(mean = mean(y3))
```

```
df %>% group_by(x4) %>% summarise(mean = mean(y4))
```

Your turn

Identify the parts that change.

Give them names.

Make a function.

Why doesn't it work?

Let's try with some dplyr code

```
df %>% group_by(x1) %>% summarise(mean = mean(y1))
```

```
df %>% group_by(x2) %>% summarise(mean = mean(y2))
```

```
df %>% group_by(x3) %>% summarise(mean = mean(y3))
```

```
df %>% group_by(x4) %>% summarise(mean = mean(y4))
```

First identify the parts that change

```
df %>% group_by(x1) %>% summarise(mean = mean(y1))  
df %>% group_by(x2) %>% summarise(mean = mean(y2))  
df %>% group_by(x3) %>% summarise(mean = mean(y3))  
df %>% group_by(x4) %>% summarise(mean = mean(y4))
```

Then give them names

df

group_var

summary_var

```
df %>% group_by(x1) %>% summarise(mean = mean(y1))
df %>% group_by(x2) %>% summarise(mean = mean(y2))
df %>% group_by(x3) %>% summarise(mean = mean(y3))
df %>% group_by(x4) %>% summarise(mean = mean(y4))
```

Now make a function

```
grouped_mean <- function(df, group_var, summary_var) {  
  df %>%  
    group_by(group_var) %>%  
    summarise(mean = mean(summary_var))  
}
```

It doesn't work 😭

```
grouped_mean <- function(df, group_var, summary_var) {  
  df %>%  
    group_by(group_var) %>%  
    summarise(mean = mean(summary_var))  
}
```

```
grouped_mean(mtcars, cyl, mpg)
```

```
#> Error: Column `group_var` is unknown
```

Vocabulary

We need some new vocabulary

Evaluated using usual R rules

```
(x - min(x)) / (max(x) - min(x))
```

```
mtcars %>%
```

```
  group_by(cyl) %>%
```

```
  summarise(mean = mean(mpg))
```

Automatically **quoted** and
evaluated in a “non-standard” way

You're already familiar with this idea

```
df <- data.frame(  
  y = 1,  
  var = 2  
)
```

```
df$y
```

```
var <- "y"  
df$var
```

Predict the output!

\$ automatically quotes the variable name

```
df <- data.frame(  
  y = 1,  
  var = 2  
)
```

```
df$y  
#> [1] 1
```

```
var <- "y"  
df$var  
#> [1] 2
```

If you want refer indirectly, must use `[[` instead

```
df <- data.frame(  
  y = 1,  
  var = 2  
)
```

```
var <- "y"  
df[[var]]  
#> [1] 1
```

	Quoted	Evaluated
Direct	<code>df\$<u>y</u></code>	<code>???</code>
Indirect	<code>???</code>	<code>var <- "y"</code> <code>df[[<u>var</u>]]</code>

	Quoted	Evaluated
Direct	<code>df\$<u>y</u></code>	<code>df[["y"]]</code>
Indirect	<code>???</code>	<code>var <- "y"</code> <code>df[[var]]</code>

	Quoted	Evaluated
Direct	<code>df\$<u>y</u></code>	<code>df[["y"]]</code>
Indirect		<code>var <- "y"</code> <code>df[[var]]</code>

Identify which arguments are auto-quoted

```
library(MASS)
```

```
mtcars2 <- subset(mtcars, cyl == 4)
```

```
with(mtcars2, sum(vs))
```

```
sum(mtcars2$am)
```

```
rm(mtcars2)
```


Can't tell? Try running the code

```
library(MASS)
```

```
#> Works
```

```
MASS
```

```
#> Error: object 'MASS' not found
```

```
# -> The 1st argument of library() is quoted
```

Can't tell? Try running the code

```
subset(mtcars, cyl == 4)
```

```
#> Works
```

```
cyl == 4
```

```
#> Error: object 'cyl' not found
```

```
# -> The 2nd argument of subset() is quoted
```

You can now identify the quoted arguments

```
library(MASS)
```

```
mtcars2 <- subset(mtcars, cyl == 4)
```

```
with(mtcars2, sum(vs))
```

```
sum(mtcars2$am)
```

```
rm(mtcars2)
```

Base R has 3 primary ways to “unquote”

Quoted/Direct	Evaluated/Indirect
<code>df\$<u>y</u></code>	<pre>x <- "y" df[[x]]</pre>
<code>library(<u>MASS</u>)</code>	<pre>x <- "MASS" library(x, character.only = TRUE)</pre>
<code>rm(<u>mtcars</u>)</code>	<pre>x <- "mtcars" rm(list = x)</pre>



`rm(list = ls())`

<https://www.tidyverse.org/articles/2017/12/workflow-vs-script/>

Identify which arguments are auto-quoted

```
library(tidyverse)
```

```
mtcars %>% pull(am)
```

```
by_cyl <- mtcars %>%  
  group_by(cyl) %>%  
  summarise(mean = mean(mpg))
```

```
ggplot(by_cyl, aes(cyl, mpg)) +  
  geom_point()
```


Identify which arguments are auto-quoted


```
library(tidyverse)
```

```
mtcars %>% pull(am)
```

```
by_cyl <- mtcars %>%  
  group_by(cyl) %>%  
  summarise(mean = mean(mpg))
```

```
ggplot(by_cyl, aes(cyl, mpg)) +  
  geom_point()
```

	Quoted	Evaluated	Tidy
Direct	<code>df\$<u>y</u></code>	<code>df[["y"]]</code>	<code>pull(df, <u>y</u>)</code>
Indirect		<code>var <- "y"</code> <code>df[[<u>var</u>]]</code>	<code>???</code>

	Quoted	Evaluated	Tidy
Direct	<code>df\$<u>y</u></code>	<code>df[["y"]]</code>	<code>pull(df, <u>y</u>)</code>
Indirect		<code>var <- "y"</code> <code>df[[<u>var</u>]]</code>	<code>var <- quo(<u>y</u>)</code> <code>pull(df, !!<u>var</u>)</code>

Everywhere in the tidyverse uses !! to unquote

Pronounced bang-bang

```
x_var <- quo(cyl)
```

```
y_var <- quo(mpg)
```

```
by_cyl <- mtcars %>%
```

```
  group_by(!!x_var) %>%
```

```
  summarise(mean = mean(!!y_var))
```

```
ggplot(by_cyl, aes(!!x_var, !!y_var)) +
```

```
  geom_point()
```

Wrapping quoting functions

New: Identify quoted vs. evaluated arguments

```
df %>% group_by(x1) %>% summarise(mean = mean(y1))
```

```
df %>% group_by(x2) %>% summarise(mean = mean(y2))
```

```
df %>% group_by(x3) %>% summarise(mean = mean(y3))
```

```
df %>% group_by(x4) %>% summarise(mean = mean(y4))
```

New: Identify quoted vs. evaluated arguments

```
df %>% group_by(x1) %>% summarise(mean = mean(y1))
```

```
df %>% group_by(x2) %>% summarise(mean = mean(y2))
```

```
df %>% group_by(x3) %>% summarise(mean = mean(y3))
```

```
df %>% group_by(x4) %>% summarise(mean = mean(y4))
```

Then identify the parts that could change

```
df %>% group_by(x1) %>% summarise(mean = mean(y1))  
df %>% group_by(x2) %>% summarise(mean = mean(y2))  
df %>% group_by(x3) %>% summarise(mean = mean(y3))  
df %>% group_by(x4) %>% summarise(mean = mean(y4))
```

These become the function arguments

df

group_var

summary_var

```
df %>% group_by(x1) %>% summarise(mean = mean(y1))  
df %>% group_by(x2) %>% summarise(mean = mean(y2))  
df %>% group_by(x3) %>% summarise(mean = mean(y3))  
df %>% group_by(x4) %>% summarise(mean = mean(y4))
```

Next write the function template & identify quoted arguments

```
grouped_mean <- function(df, group_var, summary_var) {  
  
  df %>%  
    group_by(group_var) %>%  
    summarise(mean = mean(summary_var))  
}
```


New: Wrap every quoted argument in `enquo()`

```
grouped_mean <- function(df, group_var, summary_var) {  
  group_var <- enquo(group_var)  
  summary_var <- enquo(summary_var)  
  
  df %>%  
    group_by(group_var) %>%  
    summarise(mean = mean(summary_var))  
}
```

New: And then unquote with !!

```
grouped_mean <- function(df, group_var, summary_var) {  
  group_var <- enquo(group_var)  
  summary_var <- enquo(summary_var)  
  
  df %>%  
    group_by(!!group_var) %>%  
    summarise(mean = mean(!!summary_var))  
}
```

Use the expression stored inside the variable, not literally "summary_var"

```
grouped_mean(mtcars, cyl, mpg)
```

```
grouped_mean <- function(df, group_var, summary_var) {  
  group_var <- enquo(group_var)  
  summary_var <- enquo(summary_var)  
  
  df %>%  
    group_by(!!group_var) %>%  
    summarise(mean = mean(!!summary_var))  
}
```

```
grouped_mean(mtcars, cyl, mpg)
```

```
grouped_mean <- function(df, group_var, summary_var) {  
  group_var <- quo(cyl)  
  summary_var <- quo(mpg)  
  
  df %>%  
    group_by(!!group_var) %>%  
    summarise(mean = mean(!!summary_var))  
}
```

```
grouped_mean(mtcars, cyl, mpg)
```

```
grouped_mean <- function(df, group_var, summary_var) {
```

```
  df %>%
```

```
    group_by(cyl) %>%
```

```
    summarise(mean = mean(mpg))
```

```
}
```

Is it worth it?

1:23 PM - 1:43 PM

Friday

Session 5 / programming / Lazy evaluation

The "tidy eval" framework is implemented in the rlang package and is rolling out in packages across the tidyverse and beyond. There is a lively conversation these days, as people come to terms with tidy eval and share their struggles and successes with the community. Why is this such a big deal? For starters, never before have so many people engaged with R's lazy evaluation model and been encouraged and/or required to manipulate it. I'll cover some background fundamentals that provide the rationale for tidy eval and that equip you to get the most from other talks.

Speakers: [Jenny Bryan](#)

It saves a lot of typing

```
filter(diamonds, x > 0 & y > 0 & z > 0)
```

vs

```
diamonds[  
  diamonds$x > 0 &  
  diamonds$y > 0 &  
  diamonds$z > 0,  
]
```

It saves a lot of typing

```
filter(diamonds, x > 0 & y > 0 & z > 0)
```

vs

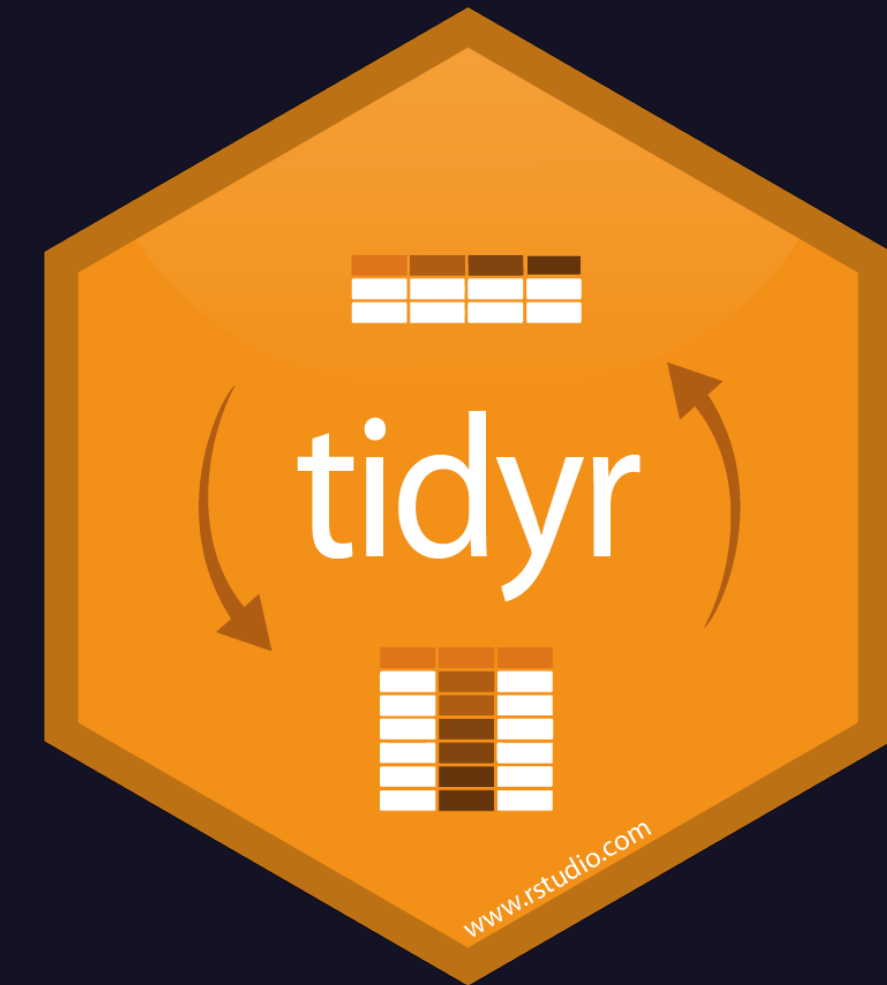
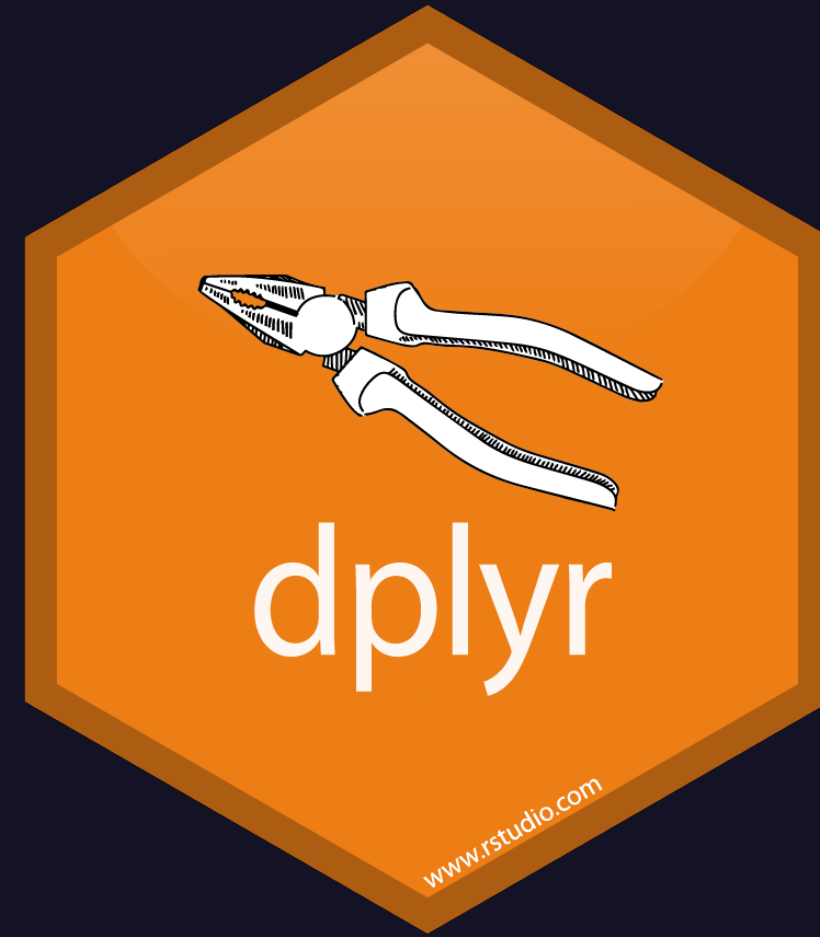
```
diamonds[  
  diamonds[["x"]] > 0 &  
  diamonds[["y"]] > 0 &  
  diamonds[["z"]] > 0,  
]
```


And makes it possible to translate to other languages

```
mtcars_db %>%  
  filter(cyl > 2) %>%  
  select(mpg:hp) %>%  
  head(10) %>%  
  show_query()
```

```
#> SELECT `mpg`, `cyl`, `disp`, `hp`  
#> FROM `mtcars`  
#> WHERE (`cyl` > 2.0)  
#> LIMIT 10
```

Tidy evaluation = principled NSE



Now for some ~~game~~ theory

1. R code is a tree
2. Unquoting builds trees
3. Environments map
names to values

Practice

Reduce the duplication here

```
df <- data.frame(  
  g = rep(c("a", "b", "c"), c(3, 2, 2)),  
  b = runif(7),  
  a = runif(7),  
  c = runif(7)  
)  
  
summarise(df, mean = mean(a), sd = sd(a), n = n())  
summarise(df, mean = mean(b), sd = sd(b), n = n())  
summarise(df, mean = mean(c), sd = sd(c), n = n())
```

```
stat_sum <- function(df, var) {  
  var <- enquo(var)  
  
  summarise(df,  
    mean = mean (!!var),  
    sd = sd (!!var),  
    n = n()  
  )  
}
```

Your turn

```
# It's often useful to compute a proportion
# of a grouped sum. Complete the function below
# to simplify this useful pattern
mtcars %>% count(cyl) %>% mutate(prop = n / sum(n))

prop <- function(df, x = n) {
  x <- enquo(x)
  ...
}
```

```
prop <- function(df, x = n) {  
  x <- enquos(x)  
  df %>% mutate(prop = !!x / sum(!!x))  
}
```


Make a reusable function for this pattern

```
counts <- starwars %>%  
  group_by(g = homeworld) %>%  
  summarise(n = n()) %>%  
  head(10) %>%  
  mutate(g = reorder(g, n))
```

```
counts %>%  
  ggplot(aes(g, n)) +  
  geom_col() +  
  coord_flip() +  
  xlab("homeworld")
```

See template on next slide

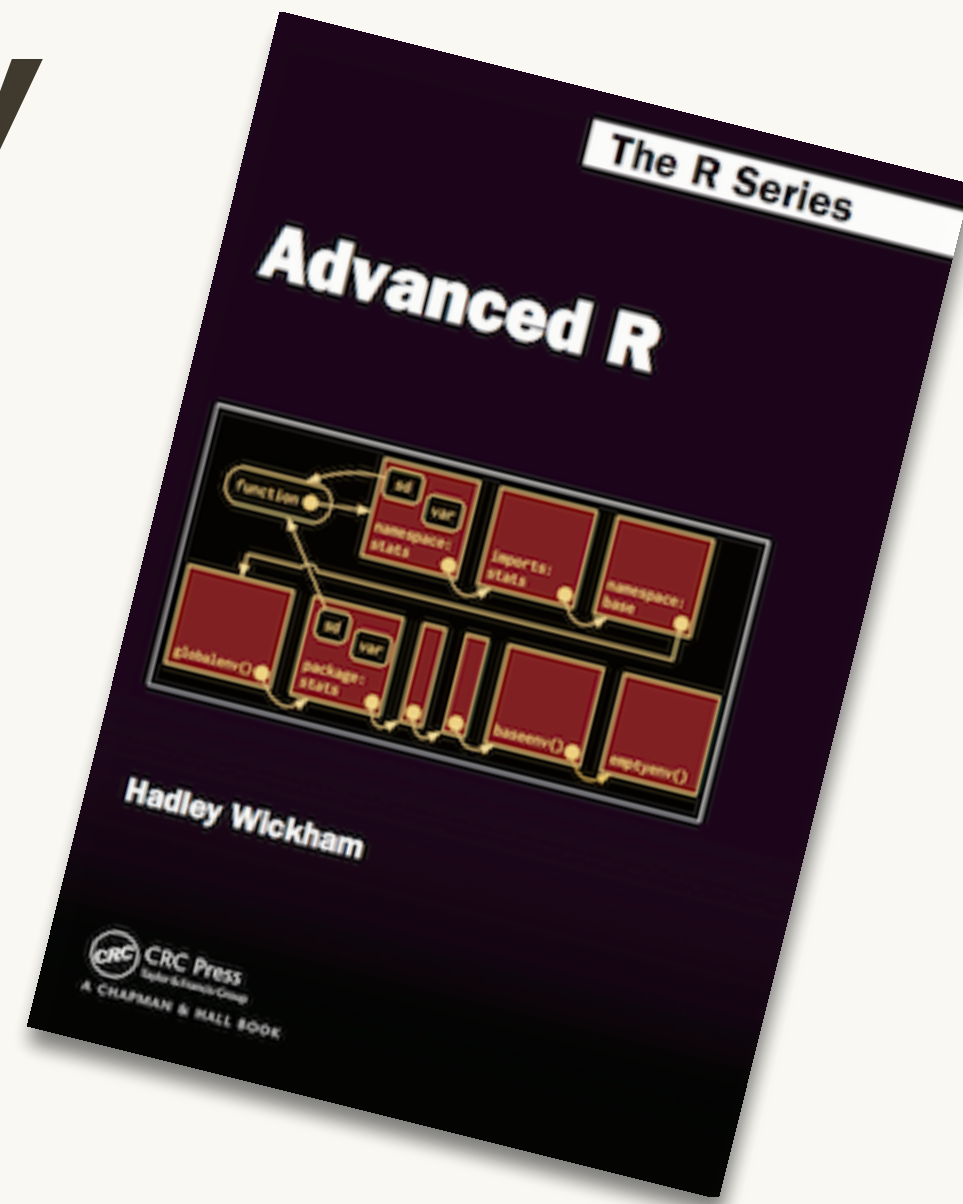
<https://twitter.com/JustTheSpring/status/1082515899821617152>

```
top_n <- function(df, x, n = 10) {  
  
}
```

```
# Challenge: can you change the basic approach  
# to better handle ties?
```

Learning more

Theory



<https://adv-r.hadley.nz/expressions.html>

<https://adv-r.hadley.nz/quasiquotation.html>

<https://adv-r.hadley.nz/evaluation.html>

<https://youtu.be/nERXS3ssntw>

Practice

<https://tidyeval.tidyverse.org>

(still a work in progress)

This work is licensed as
Creative Commons
Attribution-ShareAlike 4.0
International

To view a copy of this license, visit
<https://creativecommons.org/licenses/by-sa/4.0/>