dplyr

Hadley Wickham

Warmups

What do the following expressions return?

```
NA + 5
```

```
10 < NA
```

```
10 == NA
```

```
NA == NA
```

Why?

```
# NAs are tricky!
NA + 5
10 * NA

10 < NA
10 == NA
NA == NA
is.na(NA)
```

Name	Age	Sex
John	35	M
Mary	NA	F
Sam	NA	NA

Is Mary the same age as Jaime?
Are Sam's age and sex the same?

Studio

```
library(readr)
library(tidyr)
library(dplyr)
library(ggplot2)

library(nycflights13)

# Minor fixes for the included data
weather <- read_csv("weather.csv", col_types = list(
  date = col_datetime("%Y-%m-%d %H:%M:%S"),
  precip = col_double(),
  visib = col_double()
)) %>% mutate(
  temp = (temp - 32) * 5 / 9,
  dewp = (dewp - 32) * 5 / 9
)
```

nycflights13

- `flights` [336,776 x 16]. Every flight departing New York City in 2013.
- `weather` [26,130 x 15]
- `planes` [3,322 x 9]
- `airports` [1,397 x 7]

- **filter**: keep rows matching criteria
- **select**: pick columns by name
- **arrange**: reorder rows
- **mutate**: add new variables
- **summarise**: reduce variables to values

Structure

- First argument is a data frame
- Subsequent arguments say what to do with data frame
- Always return a data frame
- Never modify in place!

```
df <- data.frame(
  color = c("blue", "black", "blue", "blue", "black"),
  value = 1:5
)
```

df

color	value
blue	1
black	2
blue	3
blue	4
black	5

color	value
blue	1
blue	3
blue	4

```
filter(df, color == "blue")
```

df

color	value
blue	1
black	2
blue	3
blue	4
black	5

color	value
blue	1
blue	4

```
filter(df, value %in% c(1, 4))
```

	a
	b
	a | b
	a & b
	a & !b
	xor(a, b)

```
x > 1
x >= 1
x < 1
x <= 1
x != 1
x == 1
x %in% ("a", "b")
```

```
# Just prints out results
filter(flights, dest %in% c("IAH", "HOU"))
# The original is unchanged:
flights

# To create a new variable use <-
houston <- filter(flights, dest %in% c("IAH", "HOU"))
houston

# BE CAREFUL!
flights <- filter(flights, dest %in% c("IAH", "HOU"))
```