Selecting rows with `filter()`

Rhian Davies (@trianglegirl)

The Data ✈

All flights that departed from New York City in 2013.

```
library("nycflights13")
```

Typing `flights` will print out the data in our console.

```
flights
```

Using a question mark opens the help page.

```
?flights
```

Using the `View()` function opens the data in a new tab.

```
View(flights)
```

`filter()`

- Is a function in the `dplyr` package.
- Pick observations based on their values.
- Find all the flights to Hawaii.
- Find all the flights which departed on New Year's Day.
- `filter(data, condition)`

Let's try it out

Making comparisons

Symbol	Name
>	greater than
>=	greater than or equal to
<	less than
<=	less than or equal to
==	is equal to
!=	is not equal to

Careful!

```
filter(flights, month = 1)
#> `month` (`month = 1`) must not be named, do you need `==`?
```

Quiz 1

Fill in the blanks:

- Find all the flights *not* going to Atlanta (ATL).

```
filter(flights, dest ___ ___ )
```

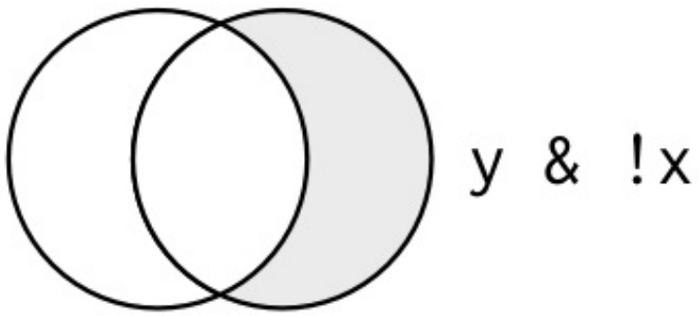
- Find all the flights which travelled more than 1500 miles.

```
filter(flights, distance ___ 1500)
```

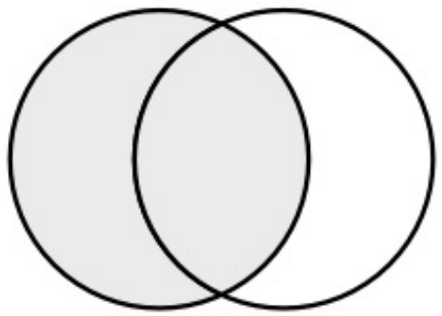
- Find all the flights to Hawaii in July.

```
filter(flights, dest ___ ___, month ___ ___)
```

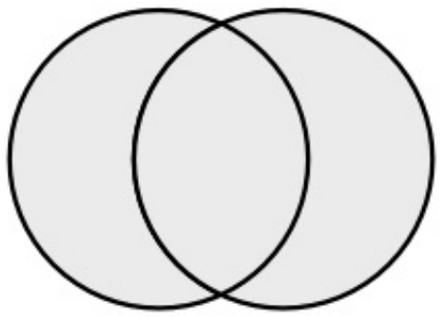
Logical operators



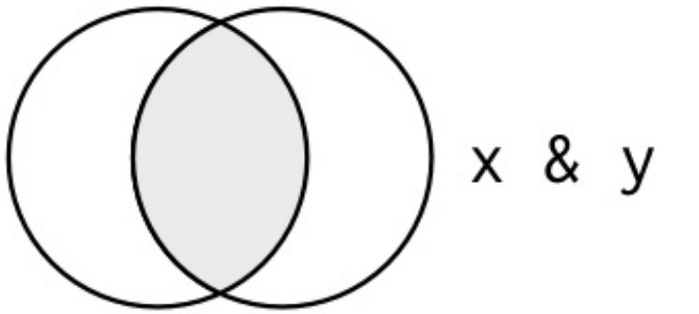
y & !x



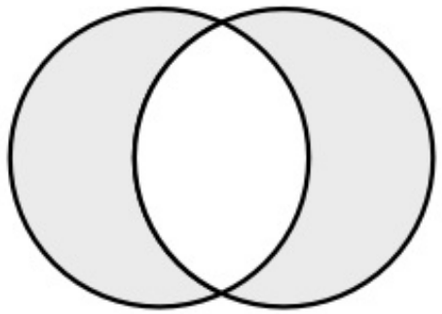
x



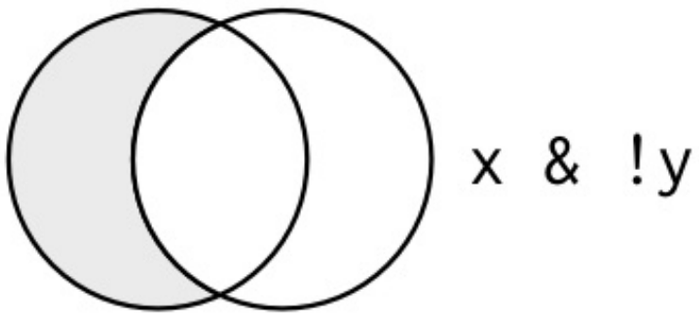
x | y



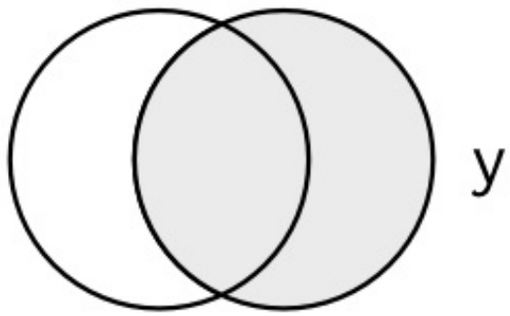
x & y



xor(x, y)



x & !y



y

Let's try it out

Quiz 2

Match the statements with the correct code. You may find drawing a Venn diagram helpful.

1. Find all United Airlines flights to Atlanta.
2. Find all United Airlines flights, not going to Atlanta.
3. Find all flights either with United Airlines or going to Atlanta (or both).
4. Find all flights going to Atlanta, not with United Airlines.

- A. `filter(flights, carrier == "UA" & dest != "ATL")`
- B. `filter(flights, carrier == "UA" | dest == "ATL")`
- C. `filter(flights, carrier == "UA" & dest == "ATL")`
- D. `filter(flights, carrier != "UA" & dest == "ATL")`

Quiz 3

Which of the statements below will *not* return all the flights occuring in Autumn?

1. `filter(flights, month >= 9 & month <= 11)`
2. `filter(flights, month > 8, month < 12 )`
3. `filter(flights, month == 9 | 10 | 11)`
4. `filter(flights, month == 9 | month == 10 | month == 11)`

Using summary statistics with filter

Remember the summary statistics that we learnt earlier?

- `mean()`
- `sd()`
- `quantile()`

Let's use those with `filter()`

Quiz 4

Fill in the gaps:

Write code to find the 5% of flights with the longest delay.

```
filter(flights, arr_delay ___ quantile(___, ___, na.rm = ___))
```

`filter()` recap

- What does it do?
- What inputs to we need?
- What does `filter()` return?
- What can help us write `filter()` code?
- What should we be wary of?