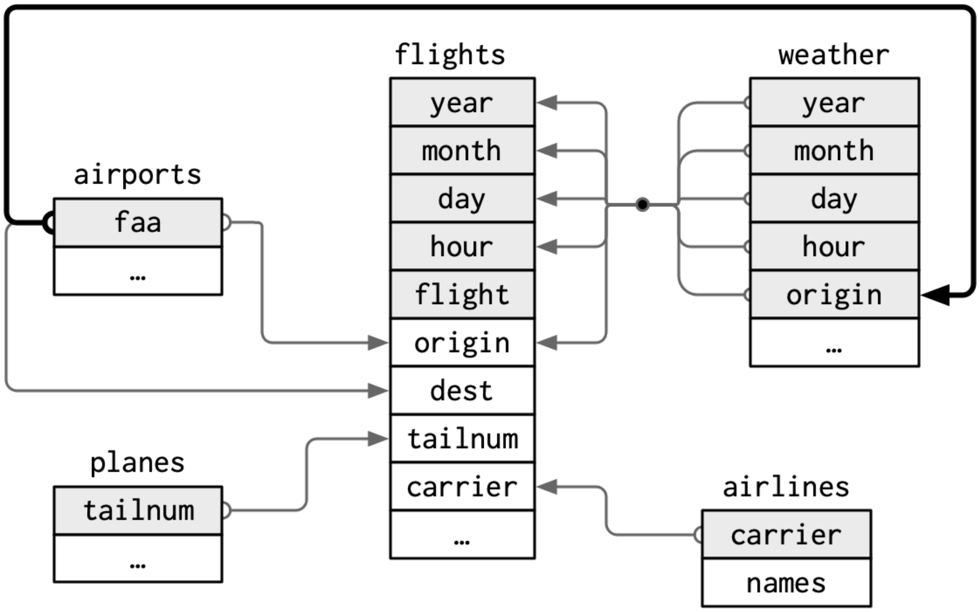
Relational data

Intro

- Many tables
- Collectively called relational data: relations are important, not just individual tables

```r
library(nycflights13)
```



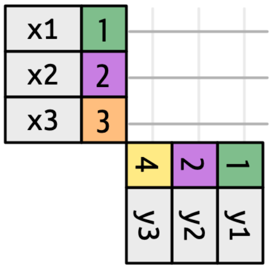
- `flights` connects to `planes` via a single variable, `tailnum`.
- `flights` connects to `airlines` through the `carrier` variable.
- `flights` connects to `airports` in two ways: via the `origin` (to `faa`) and `dest` (to `faa`) variables.
- `flights` connects to `weather` via `origin` (the location), and `year`, `month`, `day` and `hour` (the time).
- `airports` connects to `weather` through the `origin` (to `faa`) variable.

Keys

- A **key** is a variable (or set of variables) that identifies uniquely an observation in a table
  - `planes`: `tailnum`
  - `weather`: (`year`, `month`, `day`, `hour`, and `origin`)
- Primary key: uniquely identifies an observation in its own table.
  - `planes$tailnum`: uniquely identifies each plane in the planes table
- Foreign key: uniquely identifies an observation in another table.
  - `flights$tailnum` is a foreign key because it appears in the `flights` table where it matches each flight to a unique plane.

Both being primary key and foreign key possible: `origin` is part of the `weather` primary key, and is also a foreign key for the `airport` table.

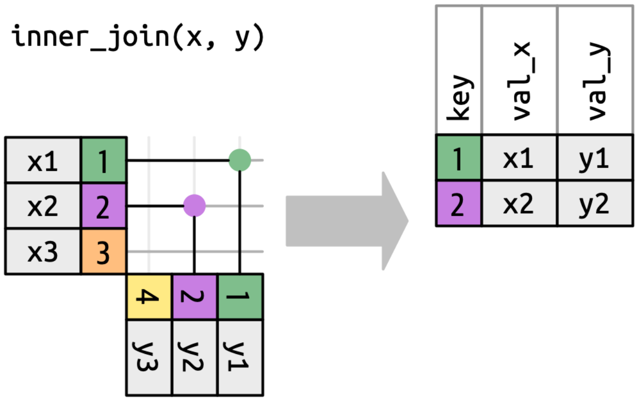
Mutating joins



- **Inner** join
- **Outer** joins
  - **Left** join
  - **Right** join
  - **Full** join

Mutating joins: Inner join

Inner join:

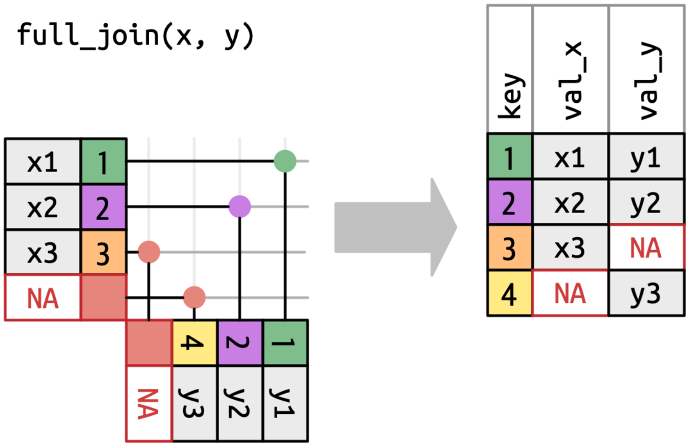


```
inner_join(x, y)
```

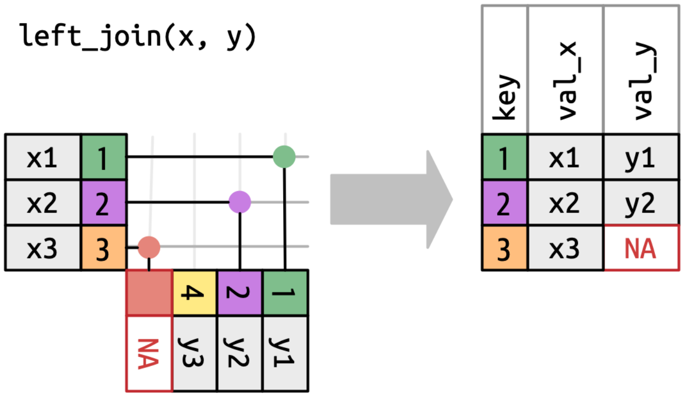
key	val_x	val_y
1	x1	y1
2	x2	y2

Mutating joins: Outer joins

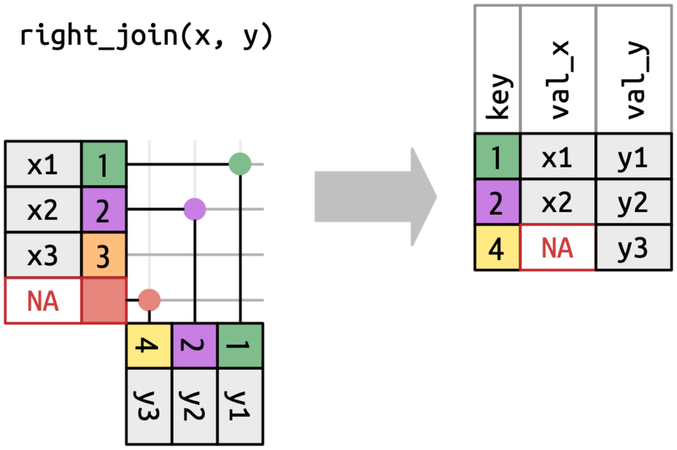
Outer joins:



key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA
4	NA	y3



key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA



key	val_x	val_y
1	x1	y1
2	x2	y2
4	NA	y3

Mutating joins: example

```
flights2 <- flights |>
  select(month, day, hour, origin, dest, tailnum, carrier)
flights2
```

```
## # A tibble: 336,776 x 7
##    month   day  hour origin dest  tailnum carrier
##    <int> <int> <dbl> <chr>  <chr> <chr>   <chr>
##  1     1     1     5 EWR    IAH    N14228  UA
##  2     1     1     5 LGA    IAH    N24211  UA
##  3     1     1     5 JFK    MIA    N619AA  AA
##  4     1     1     5 JFK    BQN    N804JB  B6
##  5     1     1     6 LGA    ATL    N668DN  DL
##  6     1     1     5 EWR    ORD    N39463  UA
##  7     1     1     6 EWR    FLL    N516JB  B6
##  8     1     1     6 LGA    IAD    N829AS  EV
##  9     1     1     6 JFK    MCO    N593JB  B6
## 10     1     1     6 LGA    ORD    N3ALAA  AA
## # i 336,766 more rows
```

```
airlines
```

```
## # A tibble: 16 x 2
##    carrier name
##    <chr>   <chr>
##  1 9E      Endeavor Air Inc.
##  2 AA      American Airlines Inc.
##  3 AS      Alaska Airlines Inc.
##  4 B6      JetBlue Airways
##  5 DL      Delta Air Lines Inc.
##  6 EV      ExpressJet Airlines Inc.
##  7 F9      Frontier Airlines Inc.
##  8 FL      AirTran Airways Corporation
##  9 HA      Hawaiian Airlines Inc.
## 10 MQ      Envoy Air
## 11 OO      SkyWest Airlines Inc.
## 12 UA      United Air Lines Inc.
## 13 US      US Airways Inc.
## 14 VX      Virgin America
## 15 WN      Southwest Airlines Co.
## 16 YV      Mesa Airlines Inc.
```

```
flights2 |>
  left_join(airlines, by = "carrier")
```

```
## # A tibble: 336,776 x 8
##    month   day  hour origin dest  tailnum carrier name
##    <int> <int> <dbl> <chr>  <chr> <chr>   <chr>   <chr>
##  1     1     1     5 EWR    IAH    N14228  UA      United Air Lines Inc.
##  2     1     1     5 LGA    IAH    N24211  UA      United Air Lines Inc.
##  3     1     1     5 JFK    MIA    N619AA  AA      American Airlines Inc.
##  4     1     1     5 JFK    BQN    N804JB  B6      JetBlue Airways
##  5     1     1     6 LGA    ATL    N668DN  DL      Delta Air Lines Inc.
##  6     1     1     5 EWR    ORD    N39463  UA      United Air Lines Inc.
##  7     1     1     6 EWR    FLL    N516JB  B6      JetBlue Airways
##  8     1     1     6 LGA    IAD    N829AS  EV      ExpressJet Airlines Inc.
##  9     1     1     6 JFK    MCO    N593JB  B6      JetBlue Airways
## 10     1     1     6 LGA    ORD    N3ALAA  AA      American Airlines Inc.
## # i 336,766 more rows
```

Mutating joins: another example

```
airports
```

```
## # A tibble: 1,458 x 8
##    faa   name                             lat    lon  alt     tz dst   tzone
##    <chr> <chr>                          <dbl>  <dbl> <dbl> <dbl> <chr> <chr>
##  1 04G   Lansdowne Airport               41.1  -80.6  1044    -5 A     America/~
##  2 06A   Moton Field Municipal Airport   32.5  -85.7   264    -6 A     America/~
##  3 06C   Schaumburg Regional             42.0  -88.1   801    -6 A     America/~
##  4 06N   Randall Airport                 41.4  -74.4   523    -5 A     America/~
##  5 09J   Jekyll Island Airport           31.1  -81.4    11    -5 A     America/~
##  6 0A9   Elizabethton Municipal Airport  36.4  -82.2  1593    -5 A     America/~
##  7 0G6   Williams County Airport         41.5  -84.5   730    -5 A     America/~
##  8 0G7   Finger Lakes Regional Airport   42.9  -76.8   492    -5 A     America/~
##  9 0P2   Shoestring Aviation Airfield    39.8  -76.6  1000    -5 U     America/~
## 10 0S9   Jefferson County Intl           48.1 -123.    108    -8 A     America/~
## # i 1,448 more rows
```

```
flights2 |>
  left_join(airports, by = c("dest" = "faa"))
```

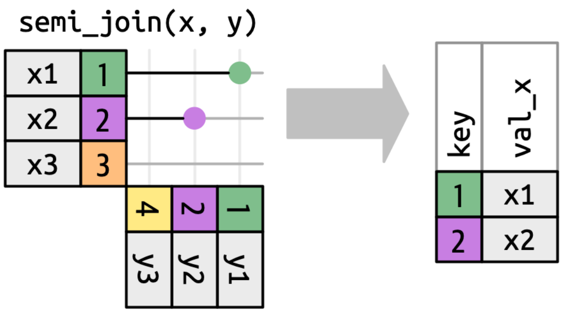
```
## # A tibble: 336,776 x 14
##    month   day  hour origin dest  tailnum carrier name     lat   lon   alt    tz
##    <int> <int> <dbl> <chr>  <chr> <chr>   <chr>   <chr>  <dbl> <dbl> <dbl> <dbl>
##  1     1     1     5 EWR    IAH   N14228  UA      Georg~  30.0 -95.3    97    -6
##  2     1     1     5 LGA    IAH   N24211  UA      Georg~  30.0 -95.3    97    -6
##  3     1     1     5 JFK    MIA   N619AA  AA      Miami~  25.8 -80.3     8    -5
##  4     1     1     5 JFK    BQN   N804JB  B6      <NA>    NA    NA      NA    NA
##  5     1     1     6 LGA    ATL   N668DN  DL      Harts~  33.6 -84.4  1026    -5
##  6     1     1     5 EWR    ORD   N39463  UA      Chica~  42.0 -87.9   668    -6
##  7     1     1     6 EWR    FLL   N516JB  B6      Fort ~  26.1 -80.2     9    -5
##  8     1     1     6 LGA    IAD   N829AS  EV      Washi~  38.9 -77.5   313    -5
##  9     1     1     6 JFK    MCO   N593JB  B6      Orlan~  28.4 -81.3    96    -5
## 10     1     1     6 LGA    ORD   N3ALAA  AA      Chica~  42.0 -87.9   668    -6
## # i 336,766 more rows
## # i 2 more variables: dst <chr>, tzone <chr>
```

```
flights2 |>
  left_join(airports, by = c("origin" = "faa"))
```

```
## # A tibble: 336,776 x 14
##    month   day  hour origin dest  tailnum carrier name     lat   lon   alt    tz
##    <int> <int> <dbl> <chr>  <chr> <chr>   <chr>   <chr>  <dbl> <dbl> <dbl> <dbl>
##  1     1     1     5 EWR    IAH   N14228  UA      Newar~  40.7 -74.2    18    -5
##  2     1     1     5 LGA    IAH   N24211  UA      La Gu~  40.8 -73.9    22    -5
##  3     1     1     5 JFK    MIA   N619AA  AA      John ~  40.6 -73.8    13    -5
##  4     1     1     5 JFK    BQN   N804JB  B6      John ~  40.6 -73.8    13    -5
##  5     1     1     6 LGA    ATL   N668DN  DL      La Gu~  40.8 -73.9    22    -5
##  6     1     1     5 EWR    ORD   N39463  UA      Newar~  40.7 -74.2    18    -5
##  7     1     1     6 EWR    FLL   N516JB  B6      Newar~  40.7 -74.2    18    -5
##  8     1     1     6 LGA    IAD   N829AS  EV      La Gu~  40.8 -73.9    22    -5
##  9     1     1     6 JFK    MCO   N593JB  B6      John ~  40.6 -73.8    13    -5
## 10     1     1     6 LGA    ORD   N3ALAA  AA      La Gu~  40.8 -73.9    22    -5
## # i 336,766 more rows
## # i 2 more variables: dst <chr>, tzone <chr>
```

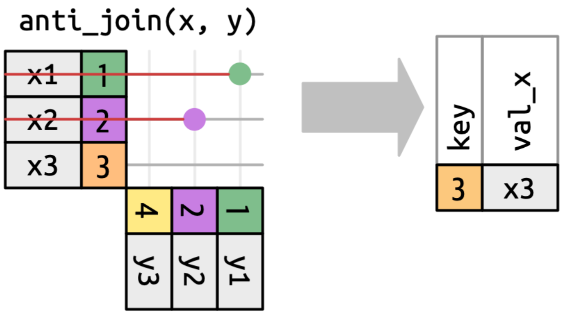
Filtering joins

- `semi_join(x, y)` **keeps** all observations in `x` that have a match in `y`.



key	val_x
1	x1
2	x2

- `anti_join(x, y)` **drops** all observations in `x` that have a match in `y`.



key	val_x
3	x3