Joins

Data stored in multiple tables

The nycflights13 package contains information on flights from NYC airports in 2013. The data is stored across several data frames:

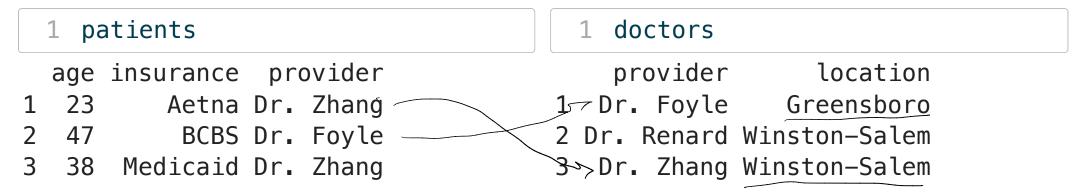
- airlines: information on each airline
- airports: information on each airport
- flights: information on each flight
- planes: information on each plane
- weather: hourly weather data

Question: What is the advantage of storing this data in multiple tables, instead of one BIG table?

Data stored in multiple tables

- Databases often contain different tables to store different information
- For example, a healthcare database could contain the following tables:
 - patients
 - doctors
 - offices
 - insurance

Joining tables



I want to add location information to the patient table. What should the resulting table look like?

age	insurance	provider	(coction
23	Aetra	Dr. Zheng	w-S
47	BCBS	Dr. Fayle	GRENDO

Left join

```
1 patients
                                       doctors
age insurance provider
                                       provider
                                                     location
                                   1 Dr. Foyle Greensboro
        Aetna Dr. Zhang
         BCBS Dr. Foyle
                                   2 Dr. Renard Winston-Salem
 38 Medicaid Dr. Zhang
                                   3 Dr. Zhang Winston-Salem
1 patients |>
    left_join(doctors, join_by(provider))
                            location
age insurance provider
                                         how to link the tables
together
        Aetna Dr. Zhang Winston-Salem
 47
         BCBS Dr. Foyle Greensboro
     Medicaid Dr. Zhang Winston-Salem
```

Left join

```
1 patients |>
2 left_join(doctors, join_by(provider))
age insurance provider location
1 23 Aetna Dr. Zhang Winston-Salem
2 47 BCBS Dr. Foyle Greensboro
3 38 Medicaid Dr. Zhang Winston-Salem
```

- Left joins are useful for adding additional information to a table
- Left joins (generally) keep the same rows as the initial dataframe (patients), and add more columns
- join_by specifies how to link the tables

Joining tables

Flights information:

```
# A tibble: 3 \times 5
  time_hour)
                     origin dest
                                  tailnum carrier
  <dttm>
                            <chr> <chr>
                                          <chr>
1 2013-01-01 05:00:00 EWR
                            IAH
                                  N14228
                                          IJA
2 2013-01-01 05:00:00 LGA IAH
                                 N24211
                                          IJΔ
3 2013-01-01 05:00:00 JFK
                         MIA
                                  N619AA
                                          AA
```

flights 1> left-join (weather) join-by (origin) tire hard)

Weather information

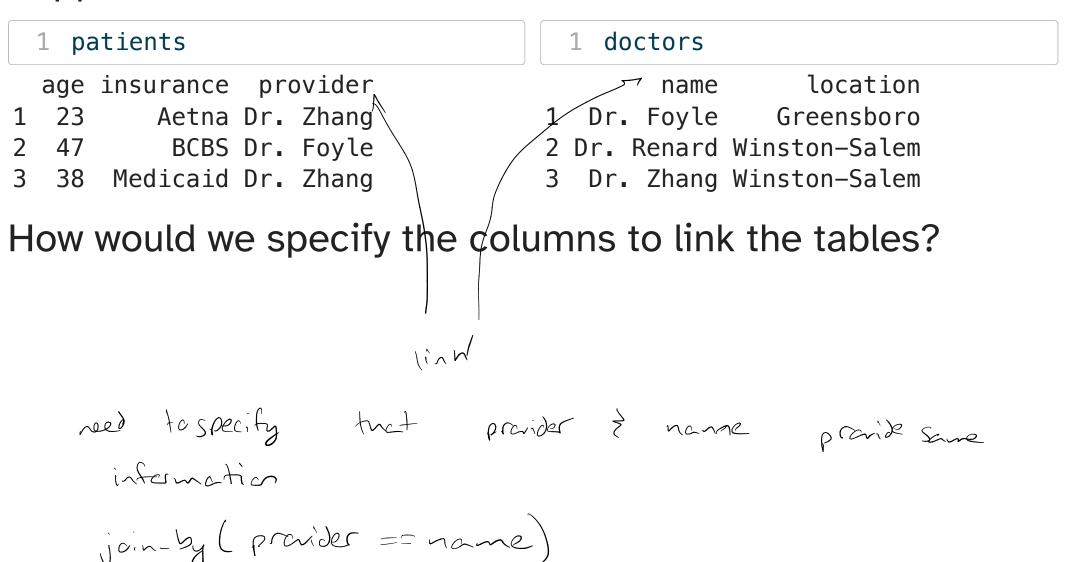
Question: What if I want to get information about the weather for each flight?

Left joins

```
flights |>
      left_join(weather, join_by(origin, time_hour))
# A tibble: 6 \times 7
  time_hour
                      origin dest tailnum carrier
                                                    temp wind_speed
  <dttm>
                      <chr>
                             <chr> <chr>
                                           <chr>
                                                   <dbl>
                                                               <dbl>
1 2013-01-01 05:00:00 EWR
                             IAH
                                   N14228
                                           IJA
                                                    39.0
                                                                12.7
2 2013-01-01 05:00:00 LGA
                                                    39.9
                                                                15.0
                             IAH
                                   N24211
                                           UA
3 2013-01-01 05:00:00 JFK
                             MIA
                                   N619AA
                                                    39.0
                                                                15.0
                                           AA
4 2013-01-01 05:00:00 JFK
                                                                15.0
                             BQN
                                   N804JB
                                           B6
                                                    39.0
5 2013-01-01 06:00:00 LGA
                             ATL
                                   N668DN
                                           DL
                                                    39.9
                                                                16.1
6 2013-01-01 05:00:00 EWR
                             ORD
                                   N39463
                                                    39.0
                                                                12.7
                                           UA
```

Joining with different names

Suppose our tables looked like this:



Joining with different names

Suppose our tables looked like this:

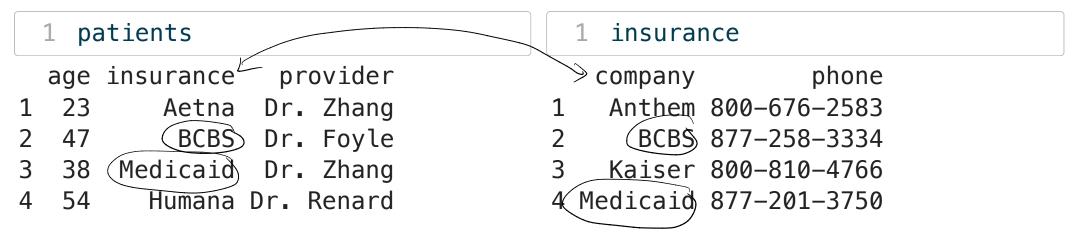
```
1 doctors
 1 patients
                                                    location
 age insurance provider
                                           name
 23 Aetna Dr. Zhang
                                   1 Dr. Foyle Greensboro
                                   2 Dr. Renard Winston-Salem
2 47 BCBS Dr. Foyle
3 38 Medicaid Dr. Zhang
                                   3 Dr. Zhang Winston-Salem
 1 patients |>
     left_join(doctors, join_by(provider == name))
 age insurance provider location
                                      provider (fran patients)
matches name (fran doctors)
         Aetna Dr. Zhang Winston-Salem
2 47 BCBS Dr. Foyle Greensboro
  38 Medicaid Dr. Zhang Winston-Salem
```

```
petients 17
leftjoin (doctor 17 select(...)) join-by(...)
```

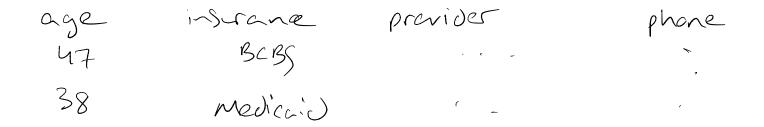
Another join

Patients in the system:

Accepted insurance:



Suppose I want insurance information only for the patients who have an accepted insurance. What should the final table look like?



Inner join

Patients in the system: Accepted insurance:

```
1 patients
                                   1 insurance
 age insurance provider
                                    company
                                                 phone
  23 Aetna Dr. Zhang
                                 1 Anthem 800-676-2583
2 47 BCBS Dr. Foyle
                                      BCBS 877-258-3334
3 38 Medicaid Dr. Zhang
                                 3 Kaiser 800-810-4766
     Humana Dr. Renard
                                 4 Medicaid 877-201-3750
 54
 1 patients |>
     inner_join(insurance, join_by(insurance == company))
 age insurance provider
                             phone
                                         lind insurance & company
         BCBS Dr. Foyle 877-258-3334
  38 Medicaid Dr. Zhang 877-201-3750
   heep only rows for which instrance / company
```

Class activity

https://sta279-f25.github.io/class_activities/ca_06.html

- Work with a neighbor on the class activity
- At the end of class, submit your work as an HTML file on Canvas (one per group, list all your names)

For next time, read:

• Chapter 26.2 in *R for Data Science*