Lecture 16: 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:

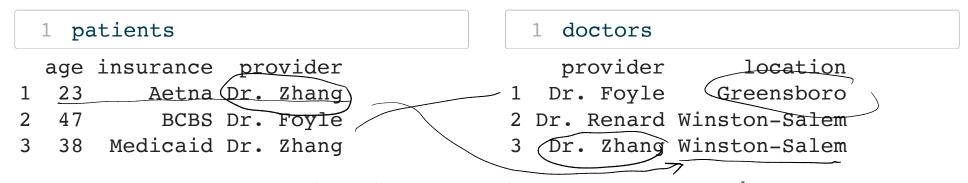
- airlines: information on each airline
- airports: information on each airport
- flights: information on each flight
- planes: information on each plane
- weather: hourly weather data

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?

Left join

```
1 patients
                                        doctors
                                        provider
 age insurance provider
                                                     location
         Aetna Dr. Zhang
 23
                                                   Greensboro
                                       Dr. Foyle
 47
          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
                                         now to link the tables together
1 23
         Aetna Dr. Zhang Winston-Salem
2 47
          BCBS Dr. Foyle Greensboro
3 38 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

Left joins in Python

Joining tables

Flights information:

Weather information

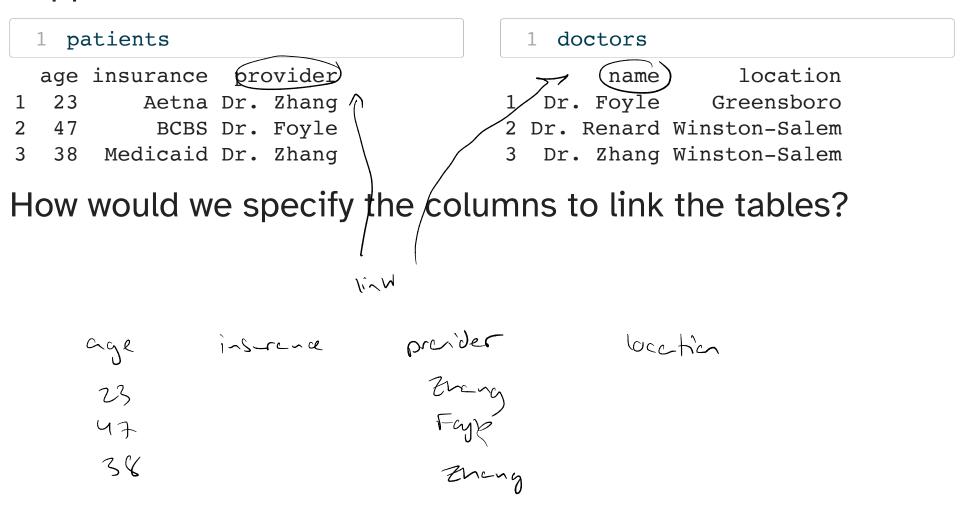
Question: What if I want to get information about the weather for each flight?

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
                       <chr>
                              <chr> <chr>
                                                      <dbl>
                                                                 <dbl>
  <dttm>
                                             <chr>
1 2013-01-01 05:00:00 EWR
                              IAH
                                    N14228
                                             UA
                                                       39.0
                                                                  12.7
2 2013-01-01 05:00:00 LGA
                                     N24211
                                                       39.9
                                                                  15.0
                              IAH
                                             UA
3 2013-01-01 05:00:00 JFK
                              MIA
                                    N619AA
                                             AA
                                                       39.0
                                                                  15.0
4 2013-01-01 05:00:00 JFK
                                                      39.0
                                                                  15.0
                                    N804JB
                                             В6
                              BQN
5 2013-01-01 06:00:00 LGA
                                                       39.9
                                                                  16.1
                              \mathsf{ATL}
                                    N668DN
                                             DL
6 2013-01-01 05:00:00 EWR
                                                       39.0
                                                                  12.7
                              ORD
                                     N39463
                                             UA
```

Joining with different names

Suppose our tables looked like this:



Joining with different names

Suppose our tables looked like this:

```
1 patients
                                         doctors
 age insurance provider
                                                       location
                                             name
 23
         Aetna Dr. Zhang
                                                     Greensboro
                                        Dr. Foyle
                                     2 Dr. Renard Winston-Salem
 47
          BCBS Dr. Foyle
3 38 Medicaid Dr. Zhang
                                     3 Dr. Zhang Winston-Salem
 1 patients |>
      left join(doctors, join by(provider == name))
                                        presider maternes name
from
from
patterts

                              location
 age insurance provider
1 23
         Aetna Dr. Zhang Winston-Salem
2 47
          BCBS Dr. Foyle Greensboro
 38 Medicaid Dr. Zhang Winston-Salem
```

In Python

```
1 pd.merge(patients, doctors, how = 'left',
2 left_on = 'provider', right_on = 'name')

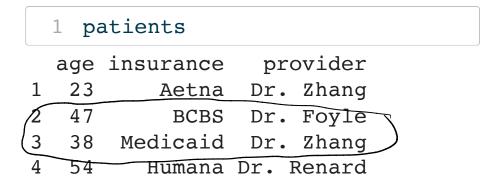
age insurance provider name location
0 23.0 Aetna Dr. Zhang Dr. Zhang Winston-Salem
1 47.0 BCBS Dr. Foyle Dr. Foyle Greensboro
2 38.0 Medicaid Dr. Zhang Dr. Zhang Winston-Salem

Pandas Heeps the Columns from the right table

(can orap the name' column after joining)
```

Another join

Patients in the system:



Accepted insurance:

```
1 insurance

company phone
1 Anthem 800-676-2583
2 BCBS 877-258-3334
3 Kaiser 800-810-4766
4 Medicaid 877-201-3750
```

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
                                        insurance
 age insurance provider
                                                     phone
                                       company
 23
                                        Anthem 800-676-2583
         Aetna Dr. Zhang
2 47
          BCBS Dr. Foyle
                                          BCBS 877-258-3334
3 38 Medicaid Dr. Zhang
                                    3 Kaiser 800-810-4766
                   vare of of
  54
        Humana Dr. Renard
                                    4 Medicaid 877-201-3750
                                      columnin "patients" of
 1 patients |>
     inner join(insurance, join by(insurance == company))
 age insurance provider
                               phone
          BCBS Dr. Foyle 877-258-3334
1 47
                                                 column in "insurance" of
2 38 Medicaid Dr. Zhang 877-201-3750
```

```
age
23
47
27
877-288---
877-201---
NA
NA
```

In Python

```
inner join
 1 pd.merge(patients, insurance, how='inner',
           left on = 'insurance', right on = 'company')
   age insurance provider company
                                        phone
0 47.0
           BCBS Dr. Foyle
                            BCBS 877-258-3334
1 38.0 Medicaid Dr. Zhang Medicaid 877-201-3750
(we could then remove "company" column if we want)
```

Class activity

https://sta279s24.github.io/class_activities/ca_lecture_16.html