

Lecture 17: Joins

STA courses next semester

- STA 214: Applied GLMs (requires MTH 111)
- STA 310: Probability (requires MTH 112)
- STA 311: Inference (requires 310)
- STA 312: Linear models (requires 310 + MTH 121 or 205)
- STA 352: Networks (requires MTH 117, 121, or 205)
- STA 362: Multivariate (requires MTH 121 or 205)
- STA 363: Stat learning (requires MTH 121 or 205)
- STA 365: Bayesian (requires 310)
- STA 368: Time series (requires 310)

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

1 patients

| | age | insurance | provider |
|---|-----|-----------|-----------|
| 1 | 23 | Aetna | Dr. Zhang |
| 2 | 47 | BCBS | Dr. Foyle |
| 3 | 38 | Medicaid | Dr. Zhang |

1 doctors

| | provider | location |
|---|------------|---------------|
| 1 | Dr. Foyle | Greensboro |
| 2 | Dr. Renard | Winston-Salem |
| 3 | Dr. Zhang | Winston-Salem |

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

| age | insurance | provider | location |
|-----|-----------|-----------|------------|
| 23 | Aetna | Dr. Zhang | W-S |
| 47 | BCBS | Dr. Foyle | Greensboro |
| 38 | Medicaid | Dr. Zhang | W-S |

Left join

```
1 patients
```

| | age | insurance | provider |
|---|-----|-----------|-----------|
| 1 | 23 | Aetna | Dr. Zhang |
| 2 | 47 | BCBS | Dr. Foyle |
| 3 | 38 | Medicaid | Dr. Zhang |

```
1 doctors
```

| | provider | location |
|---|------------|---------------|
| 1 | Dr. Foyle | Greensboro |
| 2 | Dr. Renard | Winston-Salem |
| 3 | Dr. Zhang | Winston-Salem |

```
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 |

how to link the tables together

doctors

provider
Dr. Foyle
location
G

patients

| age | insurance | provider |
|-----|-----------|-----------|
| 23 | Aetna | Dr. Zhang |
| 47 | BCBS | Dr. Foyle |

⇒

| location |
|----------|
| NA |
| G |

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

```
1 import pandas as pd
2
3 pd.merge(patients, doctors, how = 'left',
4          left_on = 'provider', right_on = 'provider')
```

← left join

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

columns to link the tables

Joining tables

Flights information:

```
# A tibble: 3 × 5
  time_hour          origin dest  tailnum carrier
  <dtm>             <chr> <chr> <chr>    <chr>
1 2013-01-01 05:00:00 EWR   IAH   N14228  UA
2 2013-01-01 05:00:00 LGA   IAH   N24211  UA
3 2013-01-01 05:00:00 JFK   MIA   N619AA  AA
```

Weather information

```
# A tibble: 3 × 4
  origin time_hour      temp wind_speed
  <chr>  <dtm>          <dbl>    <dbl>
1 EWR    2013-01-01 01:00:00  39.0     10.4
2 EWR    2013-01-01 02:00:00  39.0      8.06
3 EWR    2013-01-01 03:00:00  39.0     11.5
```

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

· left join

· join by time_hour and by origin

Left joins

```
1 flights |>
2   left_join(weather, join_by(origin, time_hour))
```

```
# A tibble: 6 × 7
```

| | time_hour <dtm> | origin <chr> | dest <chr> | tailnum <chr> | carrier <chr> | temp <dbl> | wind_speed <dbl> |
|---|---------------------|-----------------|---------------|------------------|------------------|---------------|---------------------|
| 1 | 2013-01-01 05:00:00 | EWR | IAH | N14228 | UA | 39.0 | 12.7 |
| 2 | 2013-01-01 05:00:00 | LGA | IAH | N24211 | UA | 39.9 | 15.0 |
| 3 | 2013-01-01 05:00:00 | JFK | MIA | N619AA | AA | 39.0 | 15.0 |
| 4 | 2013-01-01 05:00:00 | JFK | BQN | N804JB | B6 | 39.0 | 15.0 |
| 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 | UA | 39.0 | 12.7 |

Joining with different names

Suppose our tables looked like this:

```
1 patients
```

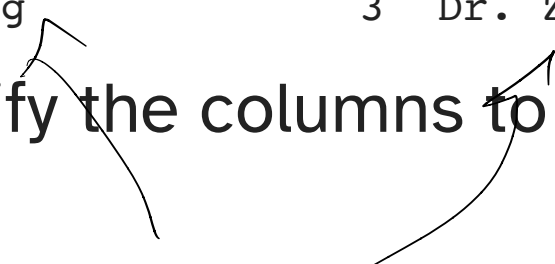
| | age | insurance | provider |
|---|-----|-----------|-----------|
| 1 | 23 | Aetna | Dr. Zhang |
| 2 | 47 | BCBS | Dr. Foyle |
| 3 | 38 | Medicaid | Dr. Zhang |

```
1 doctors
```

| | name | location |
|---|------------|---------------|
| 1 | Dr. Foyle | Greensboro |
| 2 | Dr. Renard | Winston-Salem |
| 3 | Dr. Zhang | Winston-Salem |

How would we specify the columns to link the tables?

link



Joining with different names

Suppose our tables looked like this:

```
1 patients
```

```
age insurance provider
1 23      Aetna Dr. Zhang
2 47      BCBS Dr. Foyle
3 38      Medicaid Dr. Zhang
```

```
1 doctors
```

```
name location
1 Dr. Foyle Greensboro
2 Dr. Renard Winston-Salem
3 Dr. Zhang Winston-Salem
```

```
1 patients |>
2 left_join(doctors, join_by(provider == name))
```

```
age insurance provider location
1 23      Aetna Dr. Zhang Winston-Salem
2 47      BCBS Dr. Foyle Greensboro
3 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 keeps the columns from the right table

(can drop the 'name' column after joining)

Another join

Patients in the system:

1 patients

| | age | insurance | provider |
|---|-----|-----------|------------|
| 1 | 23 | Aetna | Dr. Zhang |
| 2 | 47 | BCBS | Dr. Foyle |
| 3 | 38 | Medicaid | Dr. Zhang |
| 4 | 54 | Humana | Dr. Renard |

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?

| age | insurance | provider | phone |
|-----|-----------|-----------|--------------|
| 47 | BCBS | Dr. Foyle | 877-258-3334 |
| 38 | Medicaid | Dr. Zhang | 877-201-3750 |

Inner join

Patients in the system:

```
1 patients
```

| | age | insurance | provider |
|---|-----|-----------|------------|
| 1 | 23 | Aetna | Dr. Zhang |
| 2 | 47 | BCBS | Dr. Foyle |
| 3 | 38 | Medicaid | Dr. Zhang |
| 4 | 54 | Humana | Dr. Renard |

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 |

```
1 patients |>
```

```
2   inner_join(insurance, join_by(insurance == company))
```

| | age | insurance | provider | phone |
|---|-----|-----------|-----------|--------------|
| 1 | 47 | BCBS | Dr. Foyle | 877-258-3334 |
| 2 | 38 | Medicaid | Dr. Zhang | 877-201-3750 |

In Python

```
1 pd.merge(patients, insurance, how='inner',  
2         left_on = 'insurance', right_on = 'company')
```

inner join

| | 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 |

Class activity

<https://sta279->

[f23.github.io/class_activities/ca_lecture_17.html](https://sta279-f23.github.io/class_activities/ca_lecture_17.html)

