C: Types of joins

Key aspect 1 of joining datasets: **handling missing records**

Example: cars and gas consumption

cars:

All cars in NY

Id	State	Type
1	NY	EV
2	NY	Gas
3	NY	Gas

gas:

Gas vehicles in multiple states

Id	Gallons
2	100
3	20
4	60

Want to join using Id as the key

Two issues:

- Car 1 not in gas data
- Car 4 not in NY data

Four ways to handle discrepancies

Terminology:

cars: "left" dataset

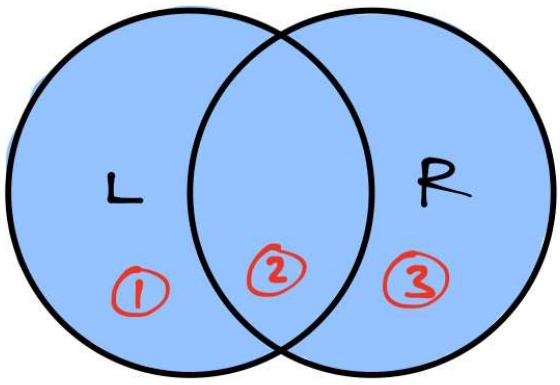
gas: "right" dataset

Case 1: outer join

Keep all records
Put in missing data as needed

Id	State	Type	Gallons
1	NY	EV	
2	NY	Gas	100
3	NY	Gas	20
4			60

Schematically: includes all keys from L **or** R



1) In L but not R:
R vars missing

2) In both:
All vars

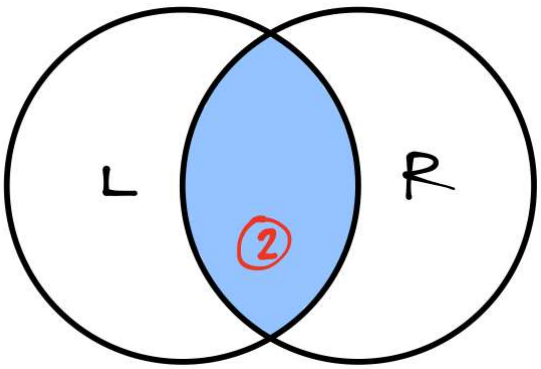
3) In R but not L:
L vars missing

Case 2: inner join

Only keep records in **both**

Id	State	Type	Gallons
2	NY	Gas	100
3	NY	Gas	20

Schematically: includes **only** keys in L **and** R



1) No records

2) All vars

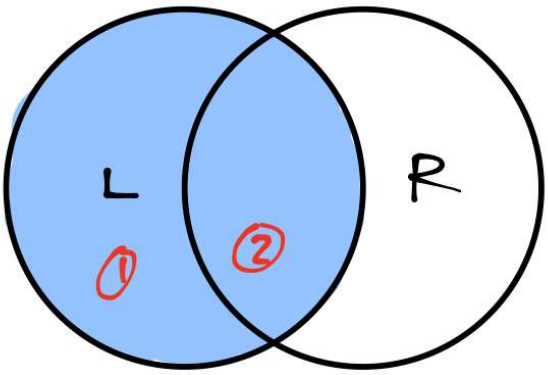
3) No records

Case 3: left join

Keep **all** records in the **left** dataset
Add missing as needed

Id	State	Type	Gallons
1	NY	EV	
2	NY	Gas	100
3	NY	Gas	20

Schematically: includes all keys in L



1) R vars missing

2) All vars

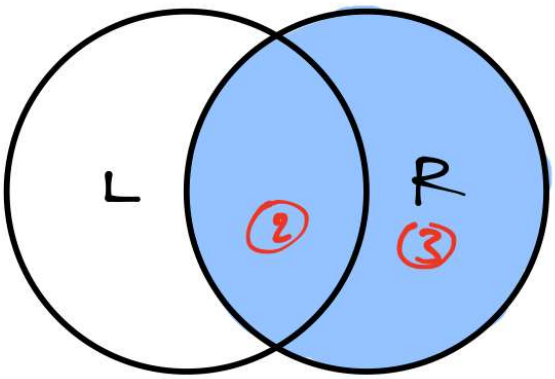
3) No records

Case 4: right join

Keep **all** records in the **right** dataset
Add missing as needed

Id	State	Type	Gallons
2	NY	Gas	100
3	NY	Gas	20
4			60

Schematically: includes all keys in R



1) No records

2) All vars

3) L vars missing

Key aspect 2 of joining datasets: **uniqueness of matches**

Three possibilities:

- One-to-one (1:1)
  Each key matches 1 key in the other dataset

  Example: joining county population and name data by FIPS code

  1. County populations, by FIPS code
  2. County names, by FIPS code

- Many-to-one (m:1) or one-to-many (1:m)
  Multiple keys in one dataset match 1 key in the other

  Example: joining multiple generators to power plants by plant code

  1. Generator data, including plant code
  2. Plant data, by plant code

     - Each plant can include many generators (m)
     - Each generator is at exactly one plant (1)

- Many-to-many (m:m)
  Multiple keys in each dataset match multiple keys in the other

  Example: authors and books

  1. Authors (may have written multiple books)
  2. Books (may have multiple authors)

  Typically handle with a junction table linking the two:

Link	Author	Title
1	William Strunk Jr	The Elements of Style
2	E.B. White	The Elements of Style
3	E.B. White	Charlotte’s Web

Examples in G14 demo.py