

## C: Joins in more detail

### Key detail #1 of joining datasets: **handling missing records**

Example: cars and gas consumption

dataset **cars**: types of cars in NY, where VIN identifies the vehicles

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

dataset **gas**: gas consumption by vehicles in multiple states

VIN	Gallons
2	100
3	20
4	60

Want to join **gas onto cars** using **VIN** as the key

Two issues:

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

**Four** ways to handle discrepancies

Terminology:

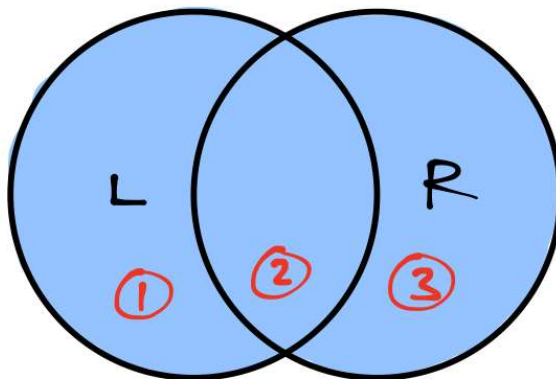
cars: left table (L)  
gas: right table (R)

### Case 1: **outer** join

- Keep **all** records
- Add missing data as needed

VIN	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**

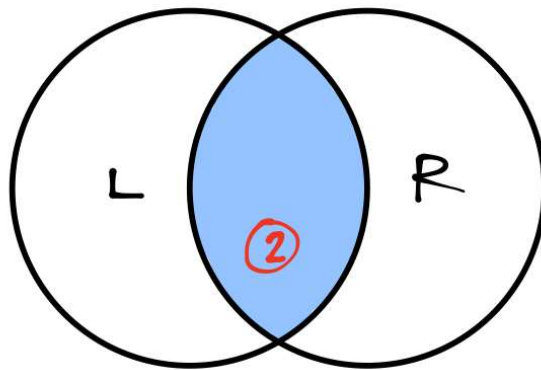
### Case 2: **inner** join

- 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

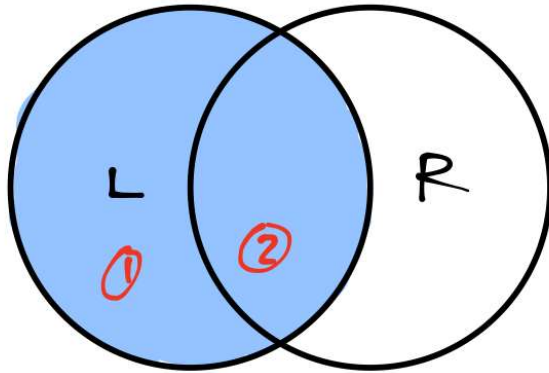
3) No records

### 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** and none not in L



1) **R vars missing**

2) All vars

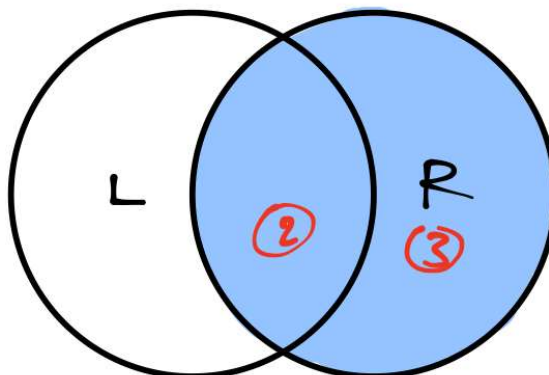
3) No records

#### 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 detail #2 of joining datasets: uniqueness of matches

Three possibilities:

### 1. One-to-one (1:1)

**Each key matches at most 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

### 2. Many-to-one (m:1) or one-to-many (1:m)

**Multiple keys** in one dataset **match 1 key** in the other

Example: joining county and state data

1. County data, including state FIPS code
2. State data, including state FIPS code
  - Each state can include many counties (m)
  - Each county is at exactly one state (1)

### 3. 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