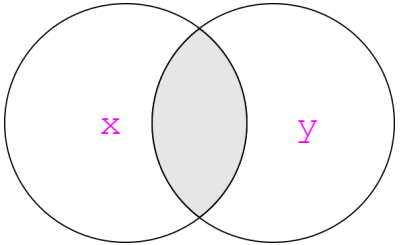
Joining Data

Getting and Cleaning Data

Inner

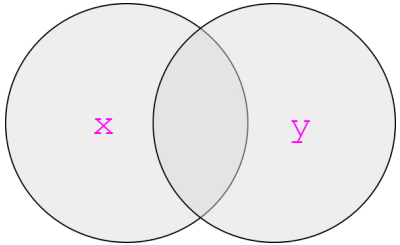
Include rows in common in both tables



`inner_join(x, y)`

Full

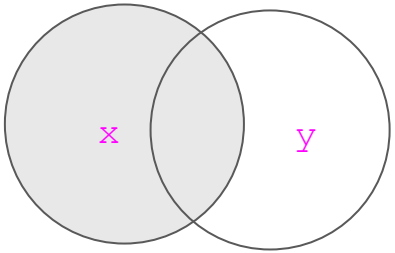
Include all rows in both tables



`full_join(x, y)`

Left

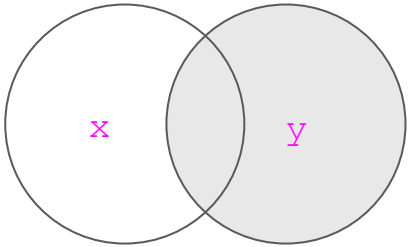
Include all rows in 1st table



`left_join(x, y)`

Right

Include all rows in 2nd table



`right_join(x, y)`

artists	
ArtistId	**Name**
1	AC/DC
3	Aerosmith
5	Alice in Chains
7	Apocalyptica
9	BackBeat
11	Black Label Society
...	...

albums		
AlbumId	**Title**	**ArtistId**
2	Balls to the Wall	2
4	Let There Be Rock	1
6	Jagged Little Pill	4
8	Warner 25 Anos	6
10	Audioslave	8
12	BackBeat Soundtrack	9
...	...	...

Inner Join - what rows have ArtistID in common in both tables?

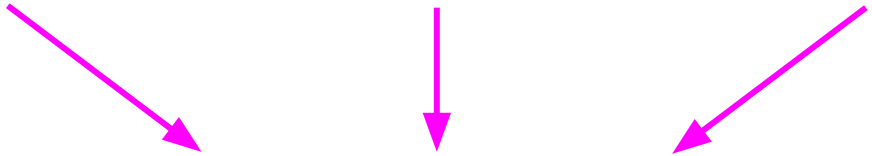
artists	
ArtistId	**Name**
1	AC/DC
3	Aerosmith
5	Alice in Chains
7	Apocalyptica
9	BackBeat
11	Black Label Society
...	...

albums		
AlbumId	**Title**	**ArtistId**
2	Balls to the Wall	2
4	Let There Be Rock	1
6	Jagged Little Pill	4
8	Warner 25 Anos	6
10	Audioslave	8
12	BackBeat Soundtrack	9
...	...	...

Inner Join: Include rows in common in both tables

artists	
ArtistId	**Name**
1	AC/DC
9	BackBeat

albums		
AlbumId	**Title**	**ArtistId**
4	Let There Be Rock	1
12	BackBeat Soundtrack	9



ArtistId	**Name**	**AlbumId**	**Title**
1	AC/DC	4	Let There Be Rock
9	BackBeat	12	BackBeat Soundtrack

```
> inner <- inner_join(artists, albums)
Joining, by = "ArtistId"
>
> ## look at output as a tibble
> as_tibble(inner)
# A tibble: 347 x 4
   ArtistId Name                 AlbumId Title
      <int> <chr>                  <int> <chr>
 1        1 AC/DC                     1 For Those About To Rock We S…
 2        2 Accept                    2 Balls to the Wall
 3        2 Accept                    3 Restless and Wild
 4        1 AC/DC                     4 Let There Be Rock
 5        3 Aerosmith                 5 Big Ones
 6        4 Alanis Morissette         6 Jagged Little Pill
 7        5 Alice In Chains           7 Facelift
 8        6 Antônio Carlos Jobim      8 Warner 25 Anos
 9        7 Apocalyptica              9 Plays Metallica By Four Cell…
10        8 Audioslave               10 Audioslave
# ... with 337 more rows
```

Left Join

artists	
ArtistId	**Name**
1	AC/DC
3	Aerosmith
5	Alice in Chains
7	Apocalyptica
9	BackBeat
11	Black Label Society
...	...

albums		
AlbumId	**Title**	**ArtistId**
2	Balls to the Wall	2
4	Let There Be Rock	1
6	Jagged Little Pill	4
8	Warner 25 Anos	6
10	Audioslave	8
12	BackBeat Soundtrack	9
...	...	...

Left Join: include all rows in first table

artists	
ArtistId	**Name**
1	AC/DC
3	Aerosmith
5	Alice in Chains
7	Apocalyptica
9	BackBeat
11	Black Label Society

albums		
AlbumId	**Title**	**ArtistId**
4	Let There Be Rock	1
12	BackBeat Soundtrack	9

ArtistId	**Name**	**AlbumId**	**Title**
1	AC/DC	4	For Those About To Rock We Salute You
3	Aerosmith	NA	NA
5	Alice in Chains	NA	NA
7	Apocalyptica	NA	NA
9	BackBeat	12	BackBeat Soundtrack
11	Black Label Society	NA	NA

```
> ## do left join
> left <- left_join(artists, albums)
Joining, by = "ArtistId"
>
> ## look at output as a tibble
> as_tibble(left)
# A tibble: 418 x 4
   ArtistId Name                 AlbumId Title
      <int> <chr>                  <int> <chr>
 1        1 AC/DC                      1 For Those About To Rock We Salute You
 2        1 AC/DC                      4 Let There Be Rock
 3        2 Accept                     2 Balls to the Wall
 4        2 Accept                     3 Restless and Wild
 5        3 Aerosmith                  5 Big Ones
 6        4 Alanis Morissette          6 Jagged Little Pill
 7        5 Alice In Chains            7 Facelift
 8        6 Antônio Carlos Jobim       8 Warner 25 Anos
 9        6 Antônio Carlos Jobim      34 Chill: Brazil (Disc 2)
10        7 Apocalyptica               9 Plays Metallica By Four Cellos
# ... with 408 more rows
```

Right Join

artists	
ArtistId	**Name**
1	AC/DC
3	Aerosmith
5	Alice in Chains
7	Apocalyptica
9	BackBeat
11	Black Label Society
...	...

albums		
AlbumId	**Title**	**ArtistId**
2	Balls to the Wall	2
4	Let There Be Rock	1
6	Jagged Little Pill	4
8	Warner 25 Anos	6
10	Audioslave	8
12	BackBeat Soundtrack	9
...	...	...

Right Join: include all rows in 2nd table

artists	
ArtistId	**Name**
1	AC/DC
9	BackBeat

albums		
AlbumId	**Title**	**ArtistId**
2	Balls to the Wall	2
4	Let There Be Rock	1
6	Jagged Little Pill	4
8	Warner 25 Anos	6
10	Audioslave	8
12	BackBeat Soundtrack	9

ArtistId	**Name**	**AlbumId**	**Title**
2	NA	2	Balls to the Wall
1	AC/DC	4	Let There Be Rock
4	NA	6	Jagged Little Pill
6	NA	8	Warner 25 Anos
8	NA	10	Audioslave
9	BackBeat	12	BackBeat Soundtrack

```
> ## do right join
> right <- right_join(as_tibble(artists), as_tibble(albums))
Joining, by = "ArtistId"
>
> ## look at output as a tibble
> as_tibble(right)
# A tibble: 347 x 4
   ArtistId Name                 AlbumId Title
      <int> <chr>                  <int> <chr>
 1        1 AC/DC                     1 For Those About To Rock We Salute You
 2        2 Accept                    2 Balls to the Wall
 3        2 Accept                    3 Restless and Wild
 4        1 AC/DC                     4 Let There Be Rock
 5        3 Aerosmith                 5 Big Ones
 6        4 Alanis Morissette         6 Jagged Little Pill
 7        5 Alice In Chains           7 Facelift
 8        6 Antônio Carlos Jobim      8 Warner 25 Anos
 9        7 Apocalyptica              9 Plays Metallica By Four Cellos
10        8 Audioslave               10 Audioslave
# ... with 337 more rows
```

Fewer columns means that there are `ArtistIds` in `artists` that are NOT in `albums`

Full Join

artists	
ArtistId	**Name**
1	AC/DC
3	Aerosmith
5	Alice in Chains
7	Apocalyptica
9	BackBeat
11	Black Label Society
...	...

albums		
AlbumId	**Title**	**ArtistId**
2	Balls to the Wall	2
4	Let There Be Rock	1
6	Jagged Little Pill	4
8	Warner 25 Anos	6
10	Audioslave	8
12	BackBeat Soundtrack	9
...	...	...

Full Join: include any row in *either* table

artists	
ArtistId	**Name**
1	AC/DC
3	Aerosmith
5	Alice in Chains
7	Apocalyptica
9	BackBeat
11	Black Label Society
...	...

albums		
AlbumId	**Title**	**ArtistId**
2	Balls to the Wall	2
4	Let There Be Rock	1
6	Jagged Little Pill	4
8	Warner 25 Anos	6
10	Audioslave	8
12	BackBeat Soundtrack	9
...	...	...

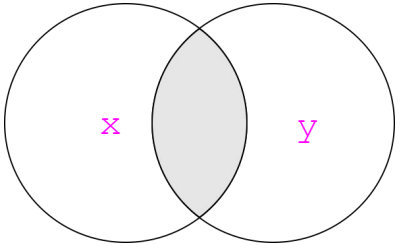
ArtistId	**Name**	**AlbumId**	**Title**
1	AC/DC	4	Let There Be Rock
2	NA	2	Balls to the Wall
3	Aerosmith	NA	NA
4	NA	6	Jagged Little Pill
5	Alice in Chains	NA	NA
6	NA	8	Warner 25 Anos
7	Apocalyptica	NA	NA
...	...	...	...

... Truncated

```
> full <- full_join(as_tibble(artists), as_tibble(albums))
Joining, by = "ArtistId"
>
> ## look at output as a tibble
> as_tibble(full)
# A tibble: 418 x 4
   ArtistId Name                 AlbumId Title
      <int> <chr>                  <int> <chr>
 1        1 AC/DC                     1 For Those About To Rock We Salute You
 2        1 AC/DC                     4 Let There Be Rock
 3        2 Accept                    2 Balls to the Wall
 4        2 Accept                    3 Restless and Wild
 5        3 Aerosmith                 5 Big Ones
 6        4 Alanis Morissette         6 Jagged Little Pill
 7        5 Alice In Chains           7 Facelift
 8        6 Antônio Carlos Jobim      8 Warner 25 Anos
 9        6 Antônio Carlos Jobim     34 Chill: Brazil (Disc 2)
10        7 Apocalyptica              9 Plays Metallica By Four Cellos
# ... with 408 more rows
```

Inner

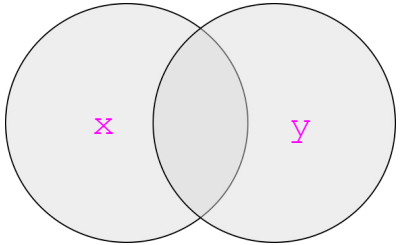
Include rows in common in both tables



`inner_join(x, y)`

Full

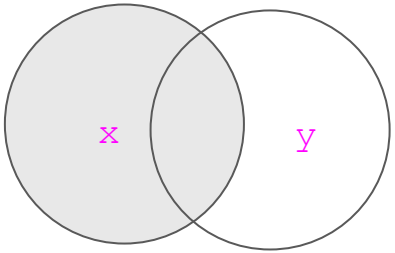
Include all rows in both tables



`full_join(x, y)`

Left

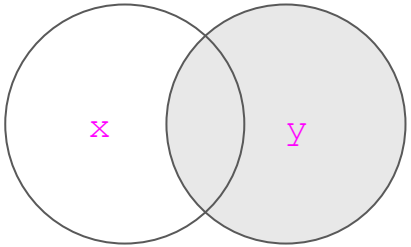
Include all rows in 1st table



`left_join(x, y)`

Right

Include all rows in 2nd table



`right_join(x, y)`

artists	
ArtistId	**Name**
1	AC/DC
3	Aerosmith
5	Alice in Chains
7	Apocalyptica
9	BackBeat
11	Black Label Society
...	...

albums		
AlbumId	**Title**	**ArtistId**
2	Balls to the Wall	2
4	Let There Be Rock	1
6	Jagged Little Pill	4
8	Warner 25 Anos	6
10	Audioslave	8
12	BackBeat Soundtrack	9
...	...	...

Semi Join: include rows in the first table that have a match in the second

artists	
ArtistId	**Name**
1	AC/DC
3	Aerosmith
5	Alice in Chains
7	Apocalyptica
9	BackBeat
11	Black Label Society

artists	
ArtistId	**Name**
1	AC/DC
9	BackBeat

! Only information from the artists table is present in the output!

```
> semi_join(artists, albums)
Joining, by = "ArtistId"
# Source:   lazy query [?? x 2]
# Database: sqlite 3.22.0 [/cloud/project/chinook.db]
   ArtistId Name
      <int> <chr>
 1        1 AC/DC
 2        2 Accept
 3        3 Aerosmith
 4        4 Alanis Morissette
 5        5 Alice In Chains
 6        6 Antônio Carlos Jobim
 7        7 Apocalyptica
 8        8 Audioslave
 9        9 BackBeat
10       10 Billy Cobham
# ... with more rows
```

Filter to only keep observations in `artists` that are also in `albums`

artists	
ArtistId	**Name**
1	AC/DC
3	Aerosmith
5	Alice in Chains
7	Apocalyptica
9	BackBeat
11	Black Label Society
...	...

albums		
AlbumId	**Title**	**ArtistId**
2	Balls to the Wall	2
4	Let There Be Rock	1
6	Jagged Little Pill	4
8	Warner 25 Anos	6
10	Audioslave	8
12	BackBeat Soundtrack	9
...	...	...

Anti Join: include rows in the first table that DO NOT have a match in the second

artists	
ArtistId	**Name**
1	AC/DC
3	Aerosmith
5	Alice in Chains
7	Apocalyptica
9	BackBeat
11	Black Label Society
...	...

artists	
ArtistId	**Name**
3	Aerosmith
5	Alice in Chains
7	Apocalyptica
11	Black Label Society

! Only information from the artists table is present in the output!

```
> anti_join(artists, albums)
Joining, by = "ArtistId"
# Source:   lazy query [?? x 2]
# Database: sqlite 3.22.0 [/cloud/project/chinook.db]
   ArtistId Name
      <int> <chr>
 1       25 Milton Nascimento & Bebeto
 2       26 Azymuth
 3       28 João Gilberto
 4       29 Bebel Gilberto
 5       30 Jorge Vercilo
 6       31 Baby Consuelo
 7       32 Ney Matogrosso
 8       33 Luiz Melodia
 9       34 Nando Reis
10       35 Pedro Luís & A Parede
# ... with more rows
```

Filter to only keep observations in `artists` that are *NOT* in `albums`

```
> glimpse(msleep)
Observations: 83
Variables: 11
$ name         <chr> "Cheetah", "Owl monkey", "M...
$ genus        <chr> "Acinonyx", "Aotus", "Aplod...
$ vore         <chr> "carni", "omni", "herbi", "...
$ order        <chr> "Carnivora", "Primates", "R...
$ conservation <chr> "lc", NA, "nt", "lc", "dome...
$ sleep_total  <dbl> 12.1, 17.0, 14.4, 14.9, 4.0...
$ sleep_rem    <dbl> NA, 1.8, 2.4, 2.3, 0.7, 2.2...
$ sleep_cycle  <dbl> NA, NA, NA, 0.1333333, 0.66...
$ awake        <dbl> 11.9, 7.0, 9.6, 9.1, 20.0, ...
$ brainwt      <dbl> NA, 0.01550, NA, 0.00029, 0...
$ bodywt       <dbl> 50.000, 0.480, 1.350, 0.019...
```

```
> conservation
# A tibble: 11 x 1
   `conservation abbreviation`
   <chr>
 1 EX = Extinct
 2 EW = Extinct in the wild
 3 CR = Critically Endangered
 4 EN = Endangered
 5 VU = Vulnerable
 6 NT = Near Threatened
 7 LC = Least Concern
 8 DD = Data deficient
 9 NE = Not evaluated
10 PE = Probably extinct (informal)
11 PEW = Probably extinct in the wild (informal)
```

The two datasets have a column in common:
conservation

```
msleep %>%
  mutate(conservation = toupper(conservation))
```

```
# A tibble: 83 x 11
   name            genus  vore  order   conservation sleep_total sleep_rem sleep_cycle awake  brainwt  bodywt
   <chr>           <chr>  <chr> <chr>   <chr>              <dbl>     <dbl>       <dbl> <dbl>    <dbl>   <dbl>
 1 Cheetah         Acino… carni Carniv… LC                  12.1      NA          NA    11.9  NA         50
 2 Owl monkey      Aotus  omni  Primat… NA                  17         1.8        NA     7    0.0155     0.48
 3 Mountain be…    Aplod… herbi Rodent… NT                  14.4       2.4        NA     9.6  NA         1.35
 4 Greater sho…    Blari… omni  Sorico… LC                  14.9       2.3         0.133  9.1  0.00029    0.019
 5 Cow             Bos    herbi Artiod… DOMESTICATED         4         0.7         0.667 20    0.423    600
 6 Three-toed …    Brady… herbi Pilosa  NA                  14.4       2.2         0.767  9.6  NA         3.85
 7 Northern fu…    Callo… carni Carniv… VU                   8.7       1.4         0.383 15.3  NA        20.5
 8 Vesper mouse    Calom… NA    Rodent… NA                   7         NA          NA    17    NA         0.045
 9 Dog             Canis  carni Carniv… DOMESTICATED        10.1       2.9         0.333 13.9  0.07      14
10 Roe deer        Capre… herbi Artiod… LC                   3         NA          NA    21    0.0982    14.8
# ... with 73 more rows
```

The values in the conservation column are now uppercase

```r
## Separate information into two columns and save to a new table
conserve <- conservation %>%
  separate(`conservation abbreviation`,
           into = c("abbreviation", "description"), sep = " = ")

## Join the two datasets together!
msleep %>%
  mutate(conservation = toupper(conservation)) %>%
  left_join(conserve, by = c("conservation" = "abbreviation"))
```

```
# A tibble: 83 x 12
   name            genus     vore  order    conservation sleep_total sleep_rem sleep_cycle awake  brainwt  bodywt description
   <chr>           <chr>     <chr> <chr>    <chr>              <dbl>     <dbl>       <dbl> <dbl>    <dbl>   <dbl> <chr>
 1 Cheetah         Acinonyx  carni Carnivo… LC                 12.1      NA          NA     11.9 NA         50     Least Conce…
 2 Owl monkey      Aotus     omni  Primates NA                 17         1.8        NA      7    0.0155    0.48   NA
 3 Mountain beav…  Aplodon…  herbi Rodentia NT                 14.4       2.4        NA      9.6 NA         1.35   Near Threat…
 4 Greater short…  Blarina   omni  Soricom… LC                 14.9       2.3         0.133  9.1  0.00029   0.019  Least Conce…
 5 Cow             Bos       herbi Artioda… DOMESTICATED        4         0.7         0.667 20    0.423   600      NA
 6 Three-toed sl…  Bradypus  herbi Pilosa   NA                 14.4       2.2         0.767  9.6 NA         3.85   NA
 7 Northern fur …  Callorh…  carni Carnivo… VU                  8.7       1.4         0.383 15.3 NA         20.5   Vulnerable
 8 Vesper mouse    Calomys   NA    Rodentia NA                  7        NA          NA     17   NA         0.045  NA
 9 Dog             Canis     carni Carnivo… DOMESTICATED       10.1       2.9         0.333 13.9  0.07      14     NA
10 Roe deer        Capreol…  herbi Artioda… LC                  3        NA          NA     21   0.0982    14.8   Least Conce…
# ... with 73 more rows
```

NAs where there is no match for the conservation status between the msleep and conservation tables

Summarizing: Joining Data

Getting and Cleaning Data