Data Science for Everyone – Data Wrangling – Joins 2

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Jordan Crouser (https://jcrouser.github.io/)

Plan for Today

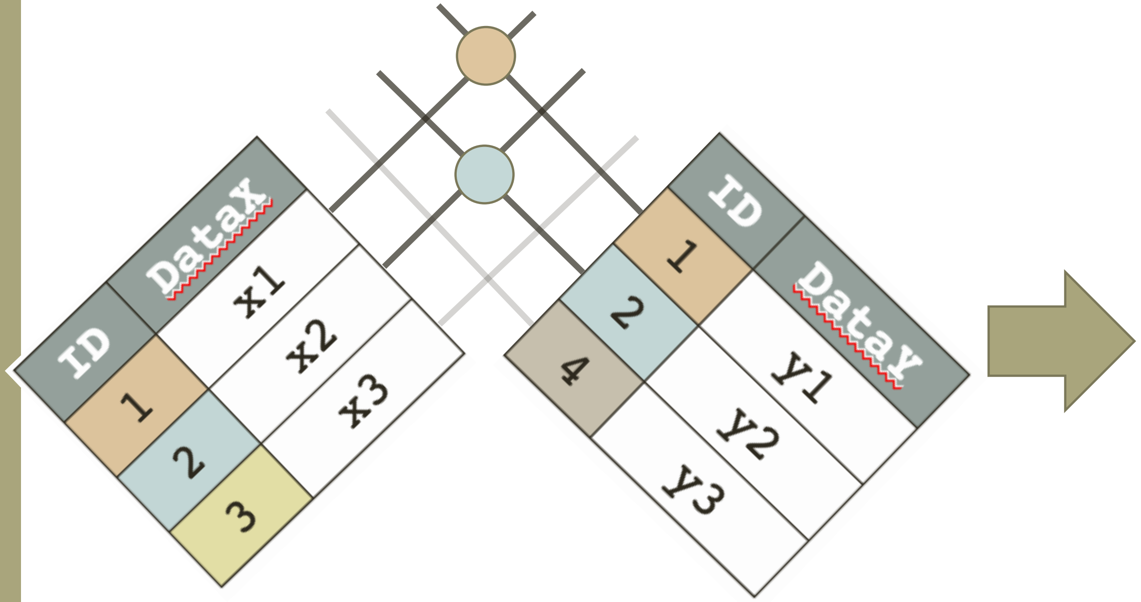
- Left and right joins
- Full joins

Joins

`inner_join()`

- Resulting table has only rows in both tables

```
Table_X %>%
   inner_join(Table_Y, by = “ID”)
```



ID	DataX
1	x1
2	x2
3	x3

ID	DataY
1	y1
2	y2
4	y3

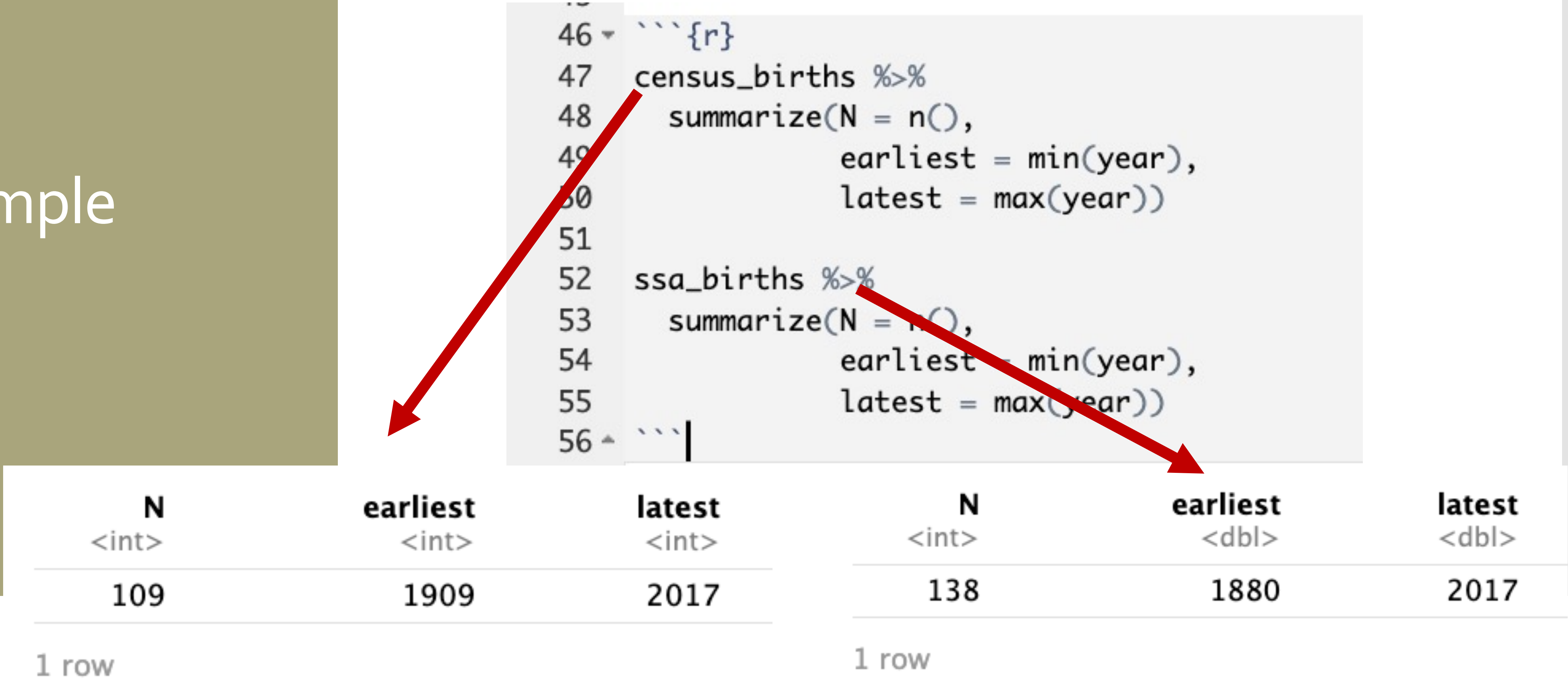
key	DataX	DataY
1	x1	y1
2	x2	y2

Example

- Let’s check that these have the same counts of babies
    - What column do SSA births and Census births share?
    - Do they have identical values in that column?

Example

- Let’s check that these have the same counts of babies
  - What column do SSA births and Census births share?
  - Do they have identical values in that column?



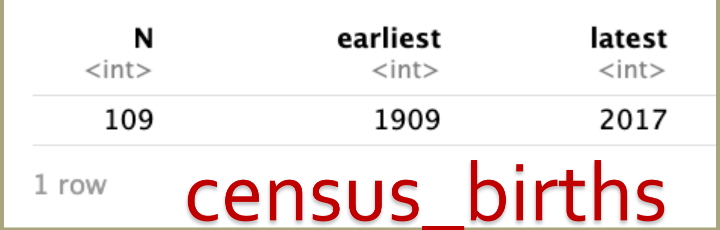
~~~
```{r}
census_births %>%
  summarize(N = n(),
            earliest = min(year),
            latest = max(year))

ssa_births %>%
  summarize(N = n(),
            earliest = min(year),
            latest = max(year))
```
~~~

```
    N       earliest    latest
<int>          <int>     <int>
  109           1909      2017
1 row
```

```
    N       earliest    latest
<int>          <dbl>     <dbl>
  138           1880      2017
1 row
```

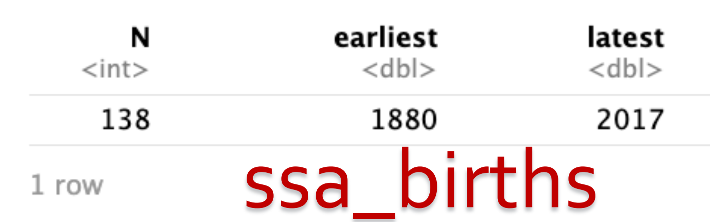
Example



N <int>	earliest <int>	latest <int>
109	1909	2017

1 row

census_births



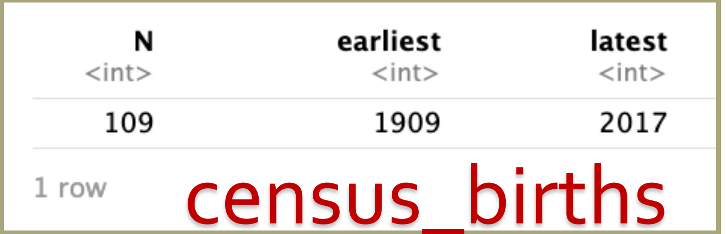
N <int>	earliest <dbl>	latest <dbl>
138	1880	2017

1 row

ssa_births

- Let’s check that these have the same counts of babies
  - What column do SSA births and Census births share?
  - Do they have identical values in that column?
  - What will happen if we do an inner join?

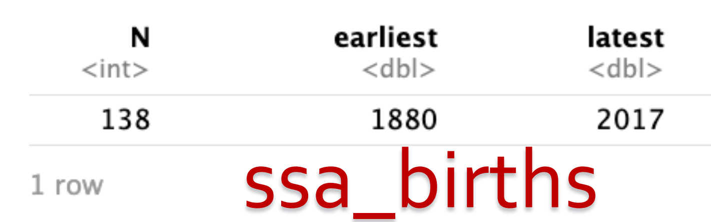
Example



N	earliest	latest
<int>	<int>	<int>
109	1909	2017

1 row

census_births

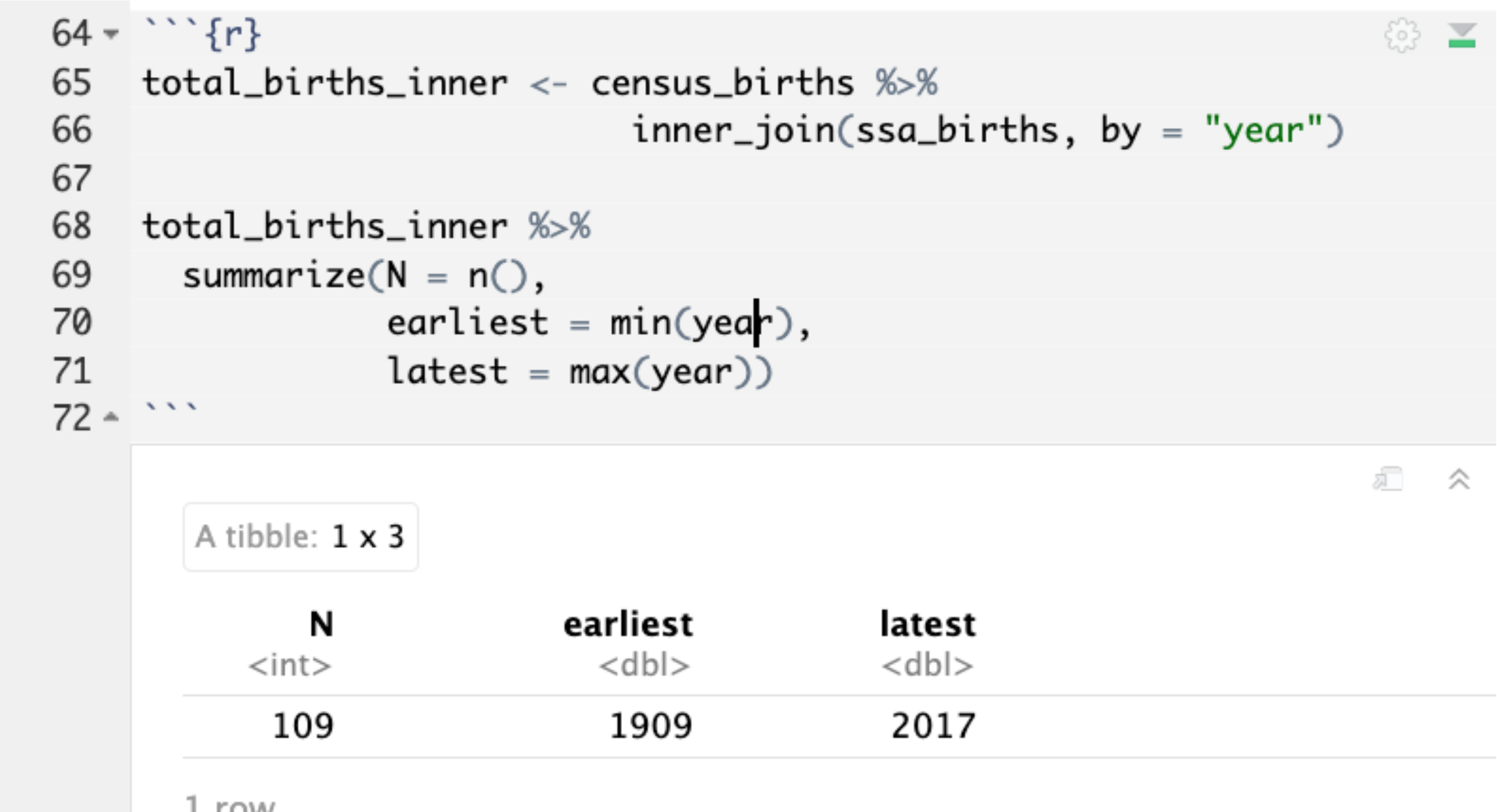


N	earliest	latest
<int>	<dbl>	<dbl>
138	1880	2017

1 row

ssa_births

- Let's check that these have the same counts of babies
  - What column do SSA births and Census births share?
  - Do they have identical values in that column?
  - What will happen if we do an inner join?



~~~
```{r}
total_births_inner <- census_births %>%
                        inner_join(ssa_births, by = "year")

total_births_inner %>%
  summarize(N = n(),
            earliest = min(year),
            latest = max(year))
```
~~~

A tibble: 1 x 3

N	earliest	latest
<int>	<dbl>	<dbl>
109	1909	2017

1 row

Joins

`left_join()`

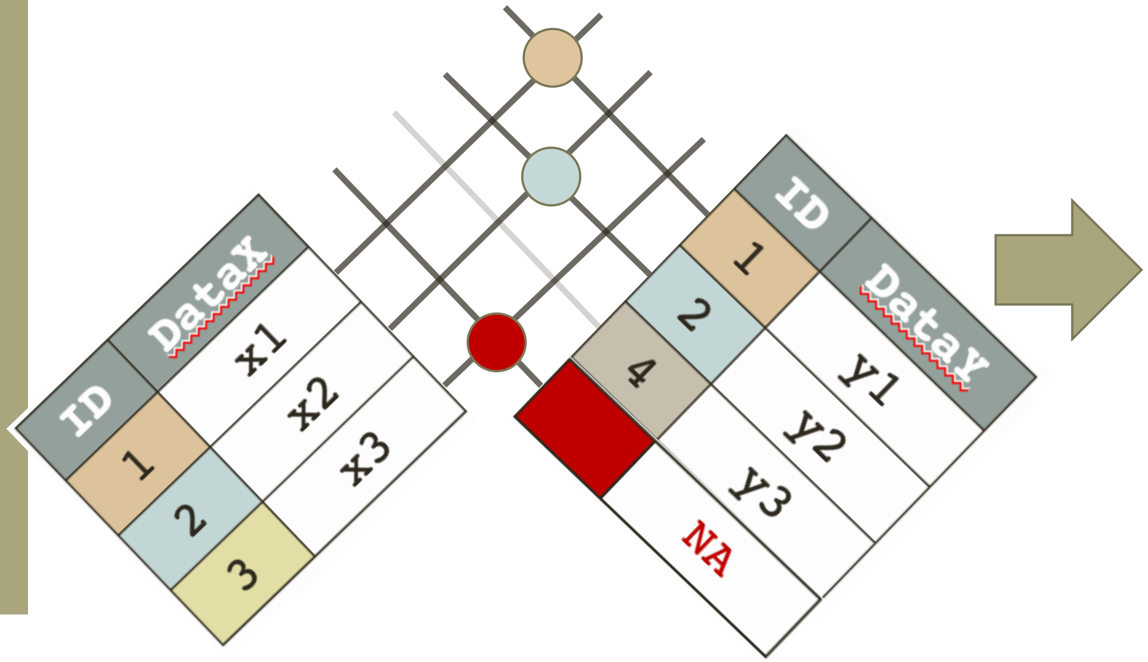
What do you think a left join does?

Joins

`left_join()`

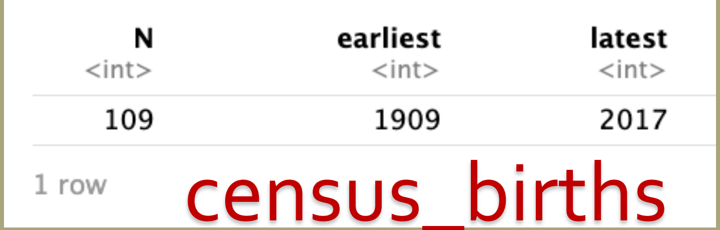
- Resulting table has all rows in left table

```
Table_X %>%
  left_join(Table_Y, by = “ID”)
```



key	DataX	DataY
1	x1	y1
2	x2	y2
3	x3	NA

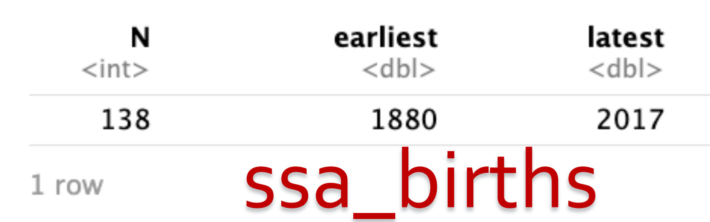
Joins



N	earliest	latest
<int>	<int>	<int>
109	1909	2017

1 row

census_births



N	earliest	latest
<int>	<dbl>	<dbl>
138	1880	2017

1 row

ssa_births

`left_join()`

- Resulting table has all rows in left table

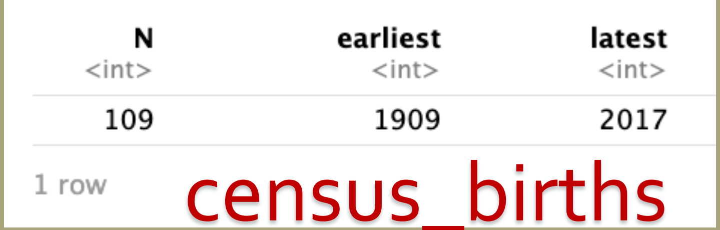
```
Table_X %>%
  left_join(Table_Y, by = "ID")
```

ID
1
2
3
x
y3
NA
3
x3
NA

How would you left join `census_births` and `ssa_births`?

What years would be in the output table?

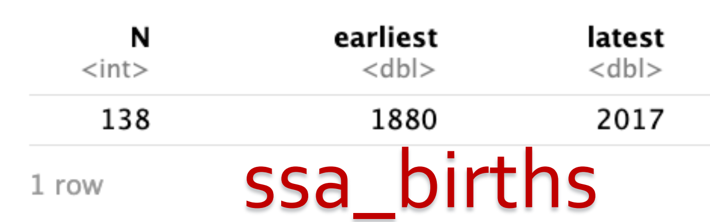
Joins



N	earliest	latest
<int>	<int>	<int>
109	1909	2017

1 row

census_births



N	earliest	latest
<int>	<dbl>	<dbl>
138	1880	2017

1 row

ssa_births

`left_join()`

- Resulting table has all rows in left table

```
census_births %>%
  left_join(ssa_births, by = “year”)
```

- Resulting table would have years 1909 - 2017

Joins

`right_join()`

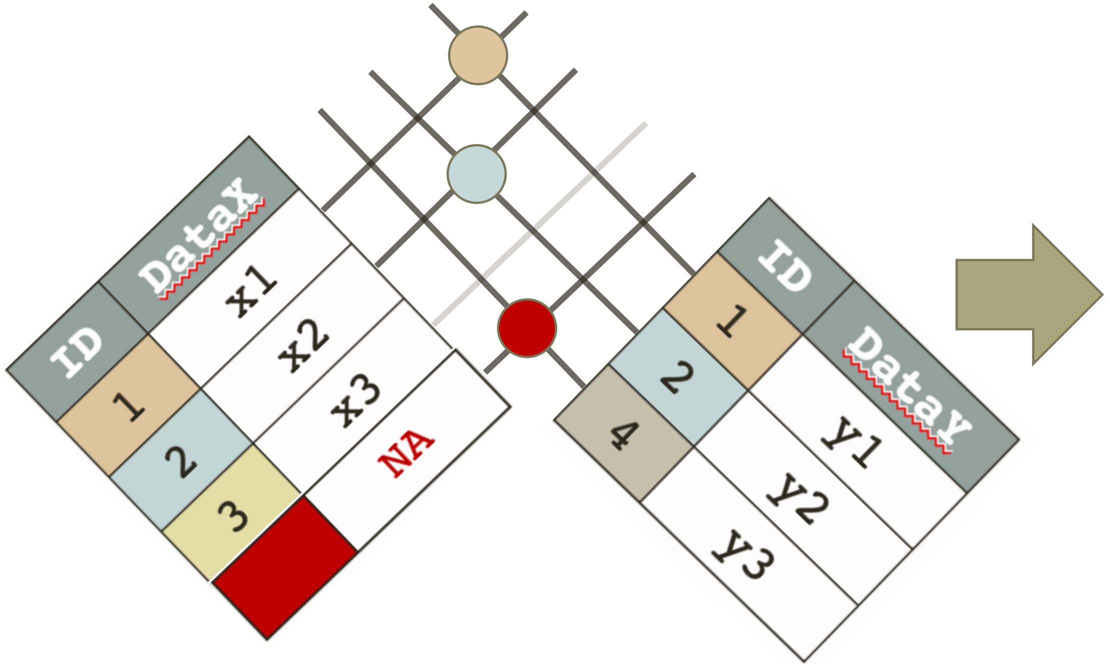
What do you think a right join does?

Joins

`right_join()`

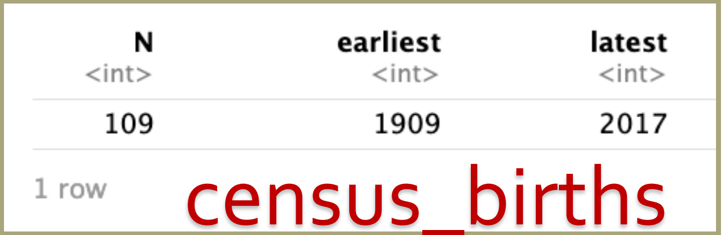
- Resulting table has all rows in right table

```
Table_X %>%
  right_join(Table_Y, by = “ID”)
```



key	DataX	DataY
1	x1	y1
2	x2	y2
4	NA	y3

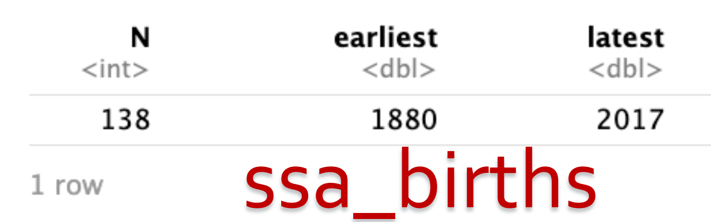
Joins



N <int>	earliest <int>	latest <int>
109	1909	2017

1 row

census_births



N <int>	earliest <dbl>	latest <dbl>
138	1880	2017

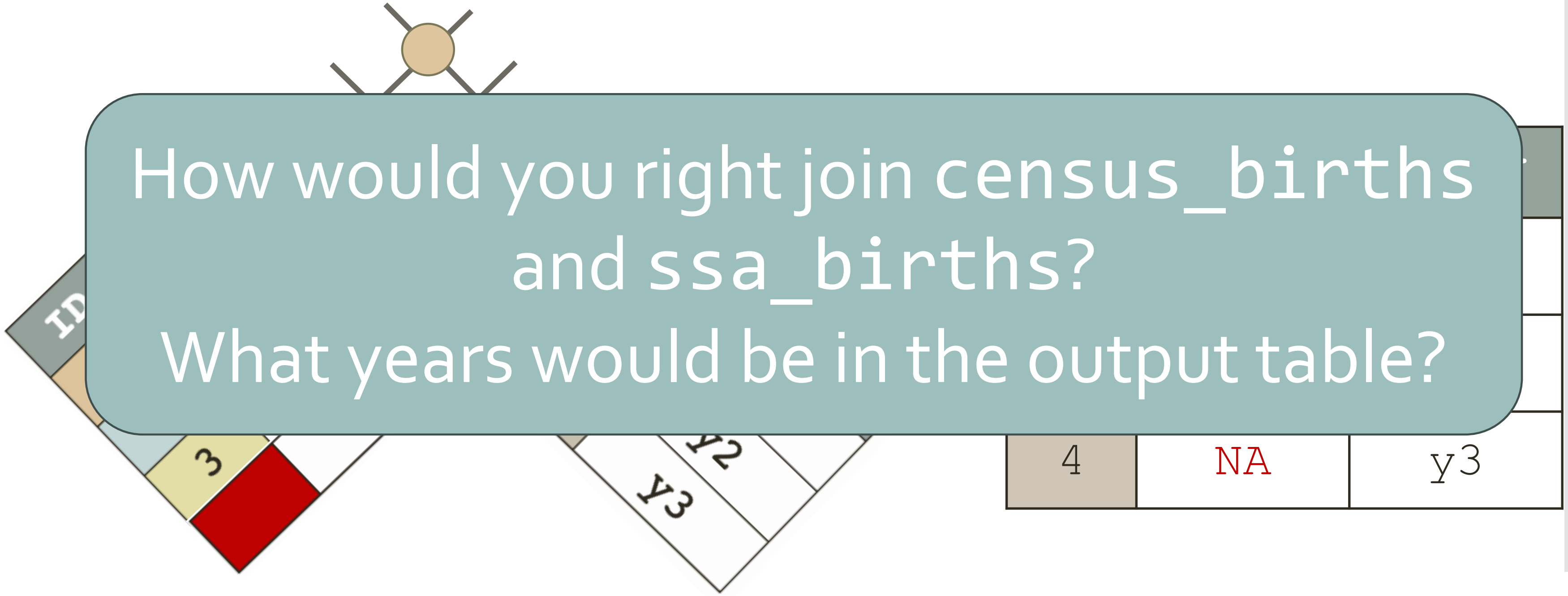
1 row

ssa_births

`right_join()`

- Resulting table has all rows in right table

```
Table_X %>%
   right_join(Table_Y, by = "ID")
```



How would you right join `census_births` and `ssa_births`?

What years would be in the output table?

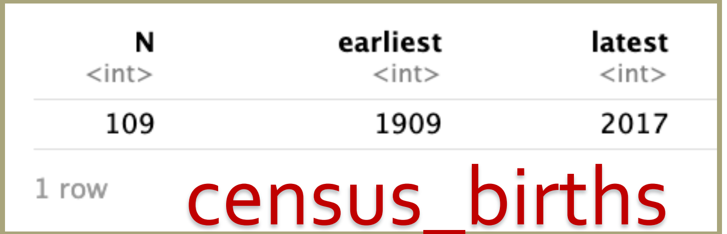
Joins

right_join()

- Resulting table has all rows in right table

```
census_births%>%
  right_join(ssa_births, by = “year”)
```

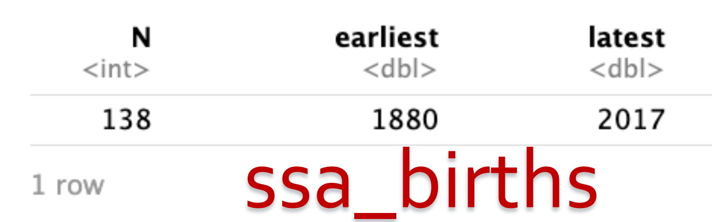
- Resulting table would have years 1880 - 2017



N <int>	earliest <int>	latest <int>
109	1909	2017

1 row

census_births

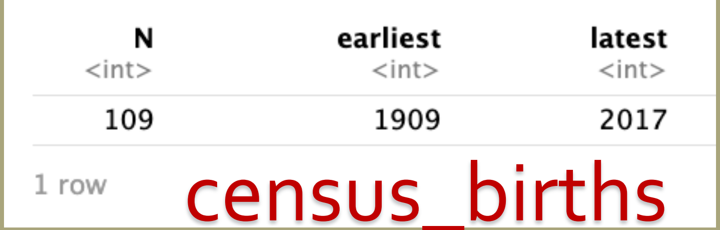


N <int>	earliest <dbl>	latest <dbl>
138	1880	2017

1 row

ssa_births

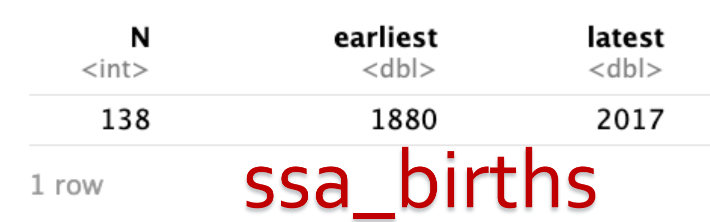
Joins



N <int>	earliest <int>	latest <int>
109	1909	2017

1 row

census_births



N <int>	earliest <dbl>	latest <dbl>
138	1880	2017

1 row

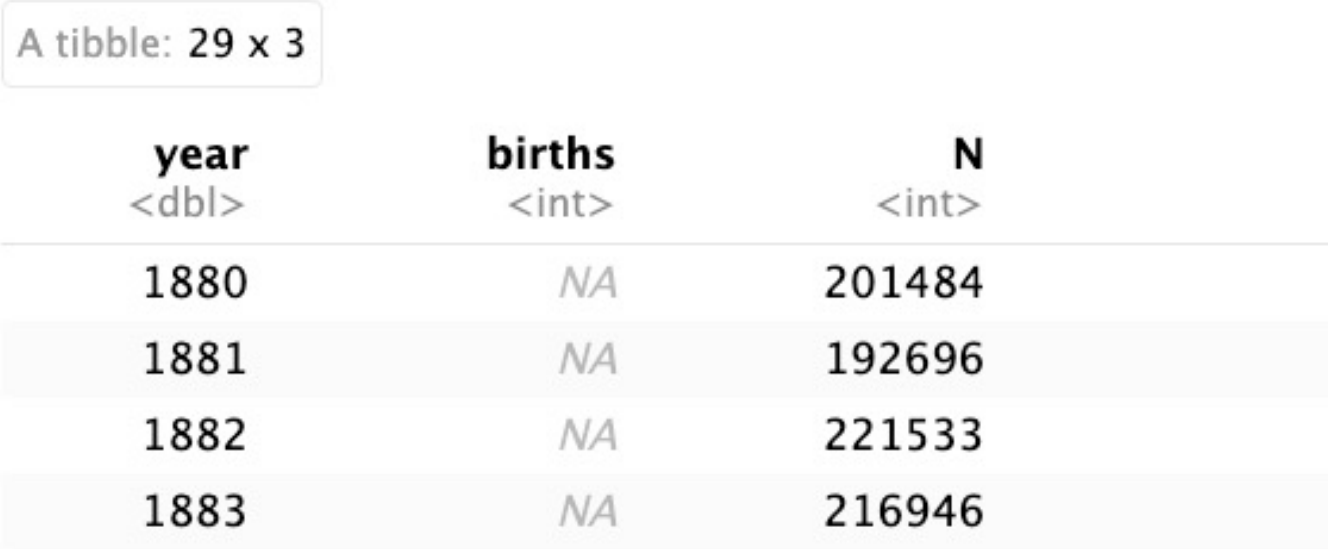
ssa_births

`right_join()`

- Resulting table has all rows in right table

```
census_births%>%
  right_join(ssa_births, by = “year”)
```

- Resulting table would have years 1880 – 2017
- Years missing in census_births would have NA data



A tibble: 29 x 3

year <dbl>	births <int>	N <int>
1880	NA	201484
1881	NA	192696
1882	NA	221533
1883	NA	216946

Joins

`full_join()`

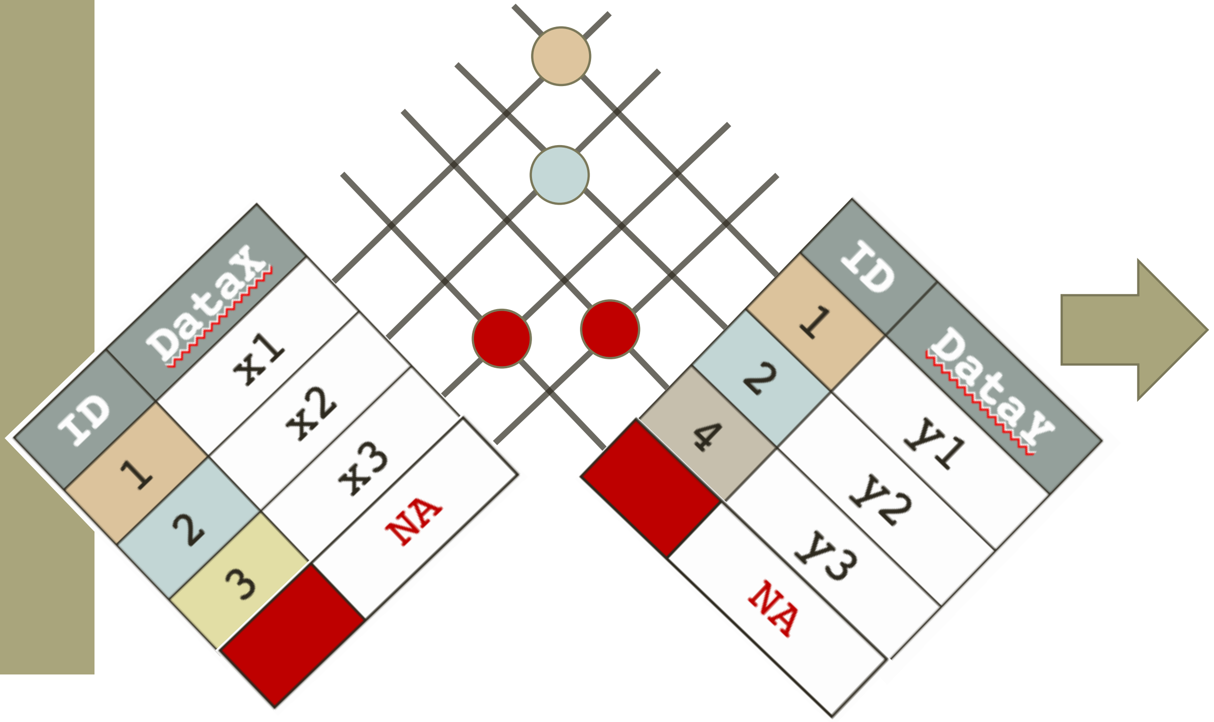
What do you think a full join does?

Joins

`full_join()`

- Resulting table has all rows in both tables

```
Table_X %>%
   full_join(Table_Y, by = "ID")
```



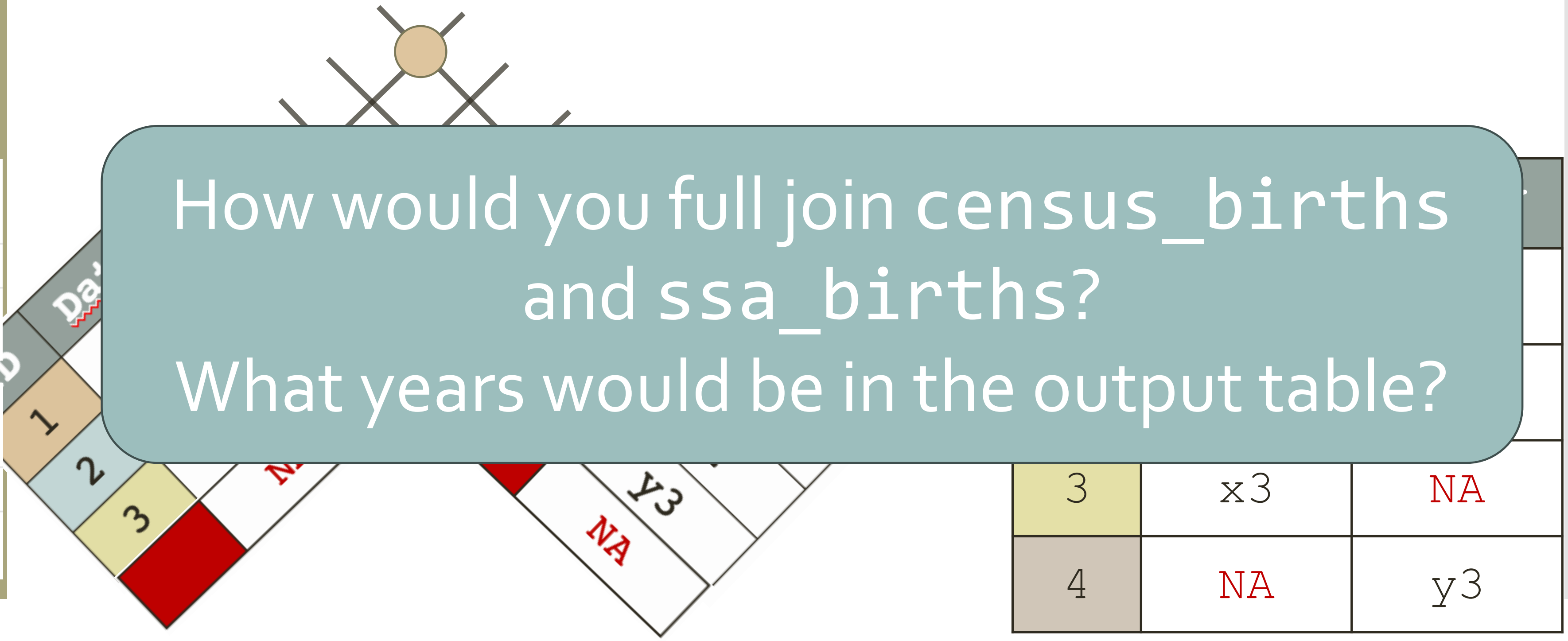
key	DataX	DataY
1	x1	y1
2	x2	y2
3	x3	NA
4	NA	y3

Joins

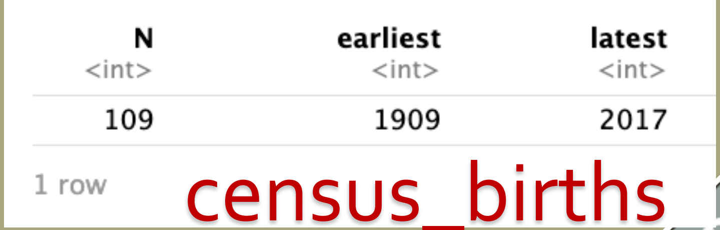
`full_join()`

- Resulting table has all rows in both tables

```
Table_X %>%
  full_join(Table_Y, by = "ID")
```

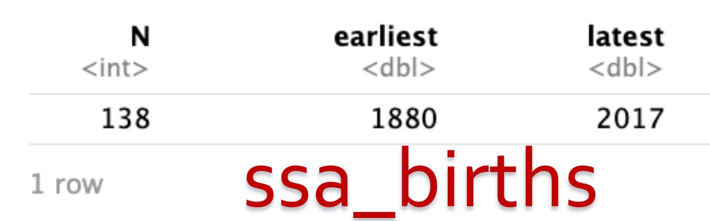


How would you full join census_births and ssa_births?
What years would be in the output table?



```
  N      earliest   latest
<int>    <int>      <int>
  109    1909       2017
1 row
```

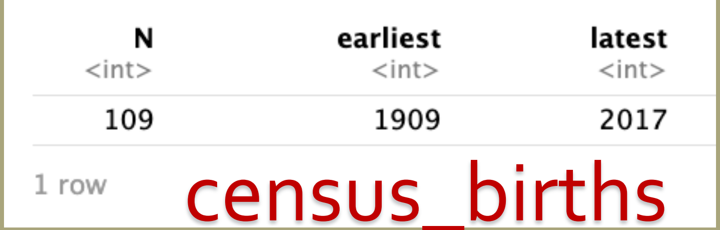
census_births



```
  N      earliest   latest
<int>    <dbl>      <dbl>
  138    1880       2017
1 row
```

ssa_births

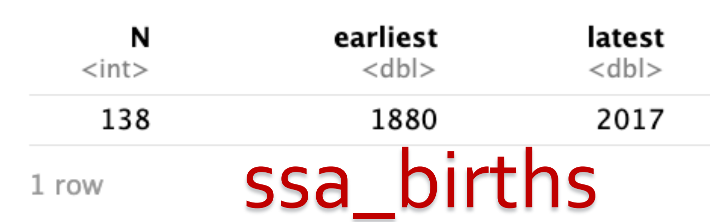
Joins



N	earliest	latest
<int>	<int>	<int>
109	1909	2017

1 row

census_births



N	earliest	latest
<int>	<dbl>	<dbl>
138	1880	2017

1 row

ssa_births

`full_join()`

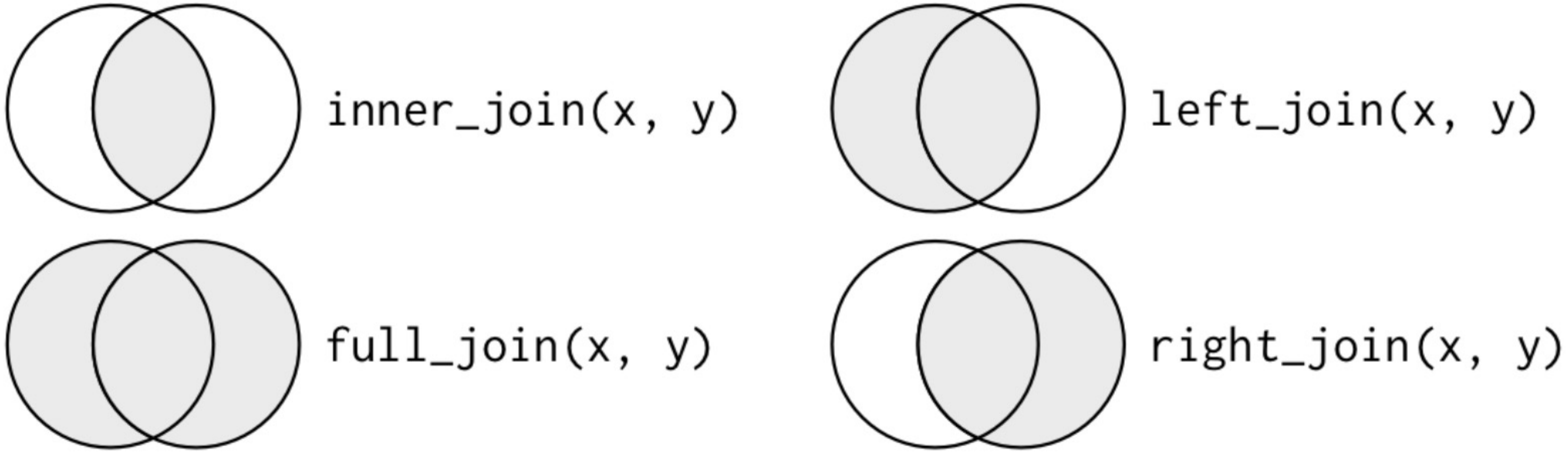
- Resulting table has all rows in both tables

```
census_births %>%
  full_join(ssa_births, by = “year”)
```

- Resulting table has years 1880 – 2017

Joins

Another way to visualize joins



Exercise

What is the difference between

left joining census_births and ssa_births

and

right joining ssa_births and census_births

?

Exercise

- Open R Studio in posit cloud
- Do a full join of ssa_births and census_births
- Add a variable indicating if the count for each year from the datasets are equal
- Are all the counts equal?
- Create a plot to compare the counts over time