Introduction to R Workshop

Session 2

Sean Nguyen

MSU BEST

Broadening Experiences in Scientific Training

Session 2: Goals

- Import external data with readr
- Introduce tidyr and dplyr
- Explore and transform data

Set working directory

- Tells R where to find your files
- Specifies where to export files
- **Windows:** shift + right click - ‘copy as path’

```
setwd(“C:\Users\(username)\Desktop”)
```

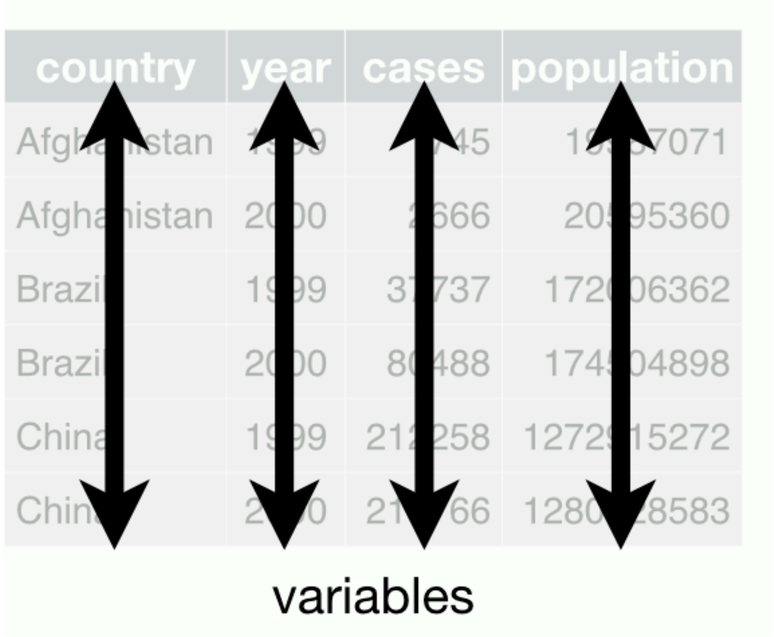
- **MacOS:** cmd + i - copy the ‘where’ path

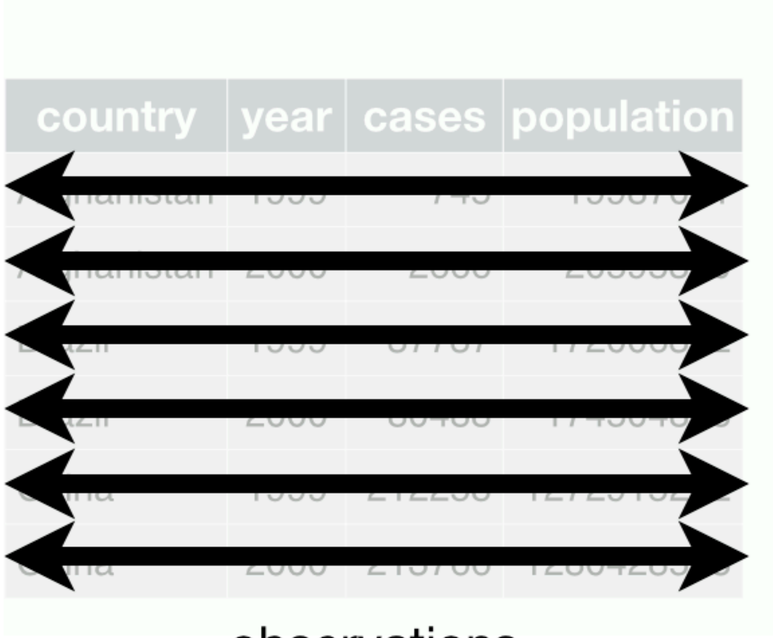
```
setwd("~/Desktop")
```

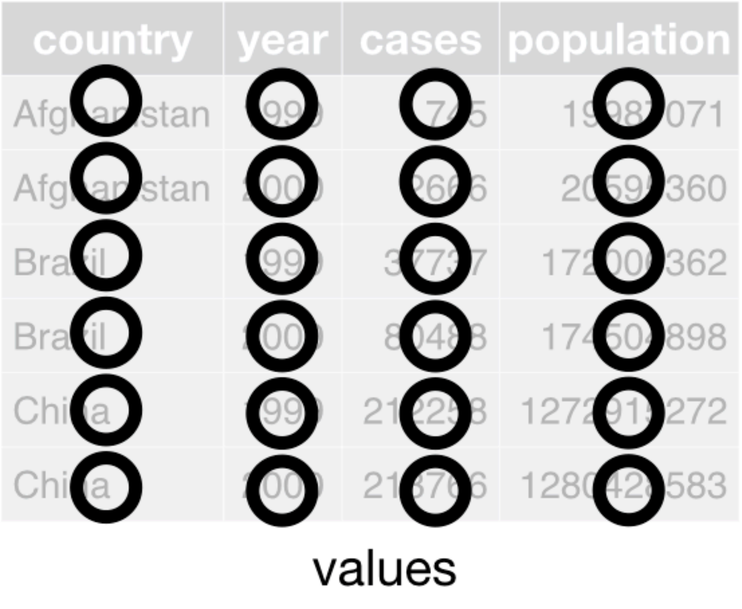
readr

- read_csv() - import .csv file
- write_csv() - export .csv file

Tidy data


country
year
cases
population
variables

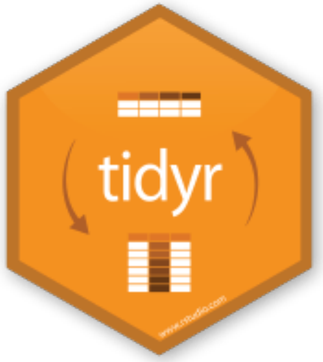

country
year
cases
population
observations


country
year
cases
population
values

tidyr

- gather() - ‘wide’ to ‘long’
- spread() - ‘long’ to ‘wide’
- separate() - split up a column
- unite() - merge multiple columns

tidyr



gather(key, time, 3:6)

gather()

- Reshapes data from ‘wide’ to ‘long’ format

Formula: gather(category, numerical, x:z)

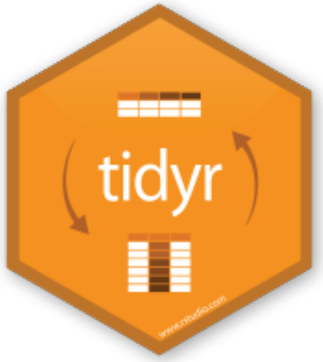
messy

id	trt	work.T1	home.T1	work.T2	home.T2
1	treatment	0.08513597	0.6158293	0.1135090	0.05190332
2	control	0.22543662	0.4296715	0.5959253	0.26417767
3	treatment	0.27453052	0.6516557	0.3580500	0.39879073
4	control	0.27230507	0.5677378	0.4288094	0.83613414

tidier

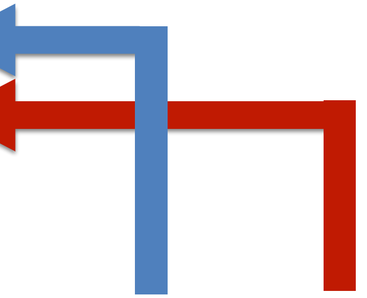
id	trt	key	time
1	treatment	work.T1	0.08513597
2	control	work.T1	0.22543662
3	treatment	work.T1	0.27453052
4	control	work.T1	0.27230507
1	treatment	home.T1	0.61582931
2	control	home.T1	0.42967153
3	treatment	home.T1	0.65165567
4	control	home.T1	0.56773775
1	treatment	work.T2	0.11350898
2	control	work.T2	0.59592531
3	treatment	work.T2	0.35804998
4	control	work.T2	0.42880942
1	treatment	home.T2	0.05190332
2	control	home.T2	0.26417767
3	treatment	home.T2	0.39879073
4	control	home.T2	0.83613414

tidyr



spread(key, time)

id	trt	work.T1	home.T1	work.T2	home.T2
1	treatment	0.08513597	0.6158293	0.1135090	0.05190332
2	control	0.22543662	0.4296715	0.5959253	0.26417767
3	treatment	0.27453052	0.6516557	0.3580500	0.39879073
4	control	0.27230507	0.5677378	0.4288094	0.83613414



spread()

- Reshapes data from ‘long’ to ‘wide’ format

Formula: spread(category, numerical)

id	trt	key	time
1	treatment	work.T1	0.08513597
2	control	work.T1	0.22543662
3	treatment	work.T1	0.27453052
4	control	work.T1	0.27230507
1	treatment	home.T1	0.61582931
2	control	home.T1	0.42967153
3	treatment	home.T1	0.65165567
4	control	home.T1	0.56773775
1	treatment	work.T2	0.11350898
2	control	work.T2	0.59592531
3	treatment	work.T2	0.35804998
4	control	work.T2	0.42880942
1	treatment	home.T2	0.05190332
2	control	home.T2	0.26417767
3	treatment	home.T2	0.39879073
4	control	home.T2	0.83613414

tidyr

separate() -split single variable into two

separate(key, into=c(“location”, ”when”), sep = “.”)

id	trt	key	time
1	treatment	work.T1	0.08513597
2	control	work.T1	0.22543662
3	treatment	work.T1	0.27453052
4	control	work.T1	0.27230507
1	treatment	home.T1	0.61582931
2	control	home.T1	0.42967153
3	treatment	home.T1	0.65165567
4	control	home.T1	0.56773775
1	treatment	work.T2	0.11350898
2	control	work.T2	0.59592531
3	treatment	work.T2	0.35804998
4	control	work.T2	0.42880942
1	treatment	home.T2	0.05190332
2	control	home.T2	0.26417767
3	treatment	home.T2	0.39879073
4	control	home.T2	0.83613414

id	trt	location	when	time
1	treatment	work	T1	0.08513597
2	control	work	T1	0.22543662
3	treatment	work	T1	0.27453052
4	control	work	T1	0.27230507
1	treatment	home	T1	0.61582931
2	control	home	T1	0.42967153
3	treatment	home	T1	0.65165567
4	control	home	T1	0.56773775
1	treatment	work	T2	0.11350898
2	control	work	T2	0.59592531
3	treatment	work	T2	0.35804998
4	control	work	T2	0.42880942
1	treatment	home	T2	0.05190332
2	control	home	T2	0.26417767
3	treatment	home	T2	0.39879073
4	control	home	T2	0.83613414

tidyr

unite() -combine two variables into one

unite(key, location, when, sep = ".")

id	trt	key	time
1	treatment	work.T1	0.08513597
2	control	work.T1	0.22543662
3	treatment	work.T1	0.27453052
4	control	work.T1	0.27230507
1	treatment	home.T1	0.61582931
2	control	home.T1	0.42967153
3	treatment	home.T1	0.65165567
4	control	home.T1	0.56773775
1	treatment	work.T2	0.11350898
2	control	work.T2	0.59592531
3	treatment	work.T2	0.35804998
4	control	work.T2	0.42880942
1	treatment	home.T2	0.05190332
2	control	home.T2	0.26417767
3	treatment	home.T2	0.39879073
4	control	home.T2	0.83613414

id	trt	location	when	time
1	treatment	work	T1	0.08513597
2	control	work	T1	0.22543662
3	treatment	work	T1	0.27453052
4	control	work	T1	0.27230507
1	treatment	home	T1	0.61582931
2	control	home	T1	0.42967153
3	treatment	home	T1	0.65165567
4	control	home	T1	0.56773775
1	treatment	work	T2	0.11350898
2	control	work	T2	0.59592531
3	treatment	work	T2	0.35804998
4	control	work	T2	0.42880942
1	treatment	home	T2	0.05190332
2	control	home	T2	0.26417767
3	treatment	home	T2	0.39879073
4	control	home	T2	0.83613414

Demo!

dplyr

- Data exploration and for transformation
- Simple syntax
- Pipeable (%>%)

Take data, then
filter on Alex, then group by sex, then
aggregate the data by adding up all
the values (by sex)

```
data %>%
    filter(name == “Alex”)%>%
    group_by(sex) %>%
    summarise( n_gender = sum(n))
```

dplyr verbs:

- **filter()** - pick specific values
- **select()** - pick specific columns
- **rename()** - change column names
- **arrange()** - sort by column values
- **mutate()** - add new columns from existing data
- **group_by()** - ‘lock-in’ by variables
- **summarise/summarize()** -aggregate data

filter()

- pick specific values
- filter(column == “value”)
- filter(column %in% c(“Mary”, ”Mari”))
- filter(column %in% c(“Mary”, ”Mari”) & year >1940)
- filter(!name == “Dave”) - filters out/omits

Try to:

- Determine flights on May 9th
- Determine flights in January and February
- Determine flights to LAX and SFO
- Determine flights delayed by >60min
- Determine flights that departed between 12am and 6am

dplyr verbs:

- **filter()** - pick specific values
- **select()** - pick specific columns
- **rename()** - change column names

select()

- pick specific columns
- select(2:49)
- select(Day, Month, Year)
- select(-xlkjgtklj) - removes “xlkjgtklj”
- select(starts_with(delay): names starts with delay

rename()

- change column names

Formula: rename(new_column = old_column)

rename(patient_ID = id,
hours = time)

id	trt	location	when	time
1	treatment	work	T1	0.08513597
2	control	work	T1	0.22543662
3	treatment	work	T1	0.27453052
4	control	work	T1	0.27230507
1	treatment	home	T1	0.61582931
2	control	home	T1	0.42967153
3	treatment	home	T1	0.65165567
4	control	home	T1	0.56773775
1	treatment	work	T2	0.11350898
2	control	work	T2	0.59592531
3	treatment	work	T2	0.35804998

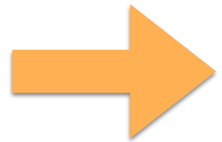
patient_ID	trt	location	when	hours
1	treatment	work	T1	0.08513597
2	control	work	T1	0.22543662
3	treatment	work	T1	0.27453052
4	control	work	T1	0.27230507
1	treatment	home	T1	0.61582931
2	control	home	T1	0.42967153
3	treatment	home	T1	0.65165567
4	control	home	T1	0.56773775
1	treatment	work	T2	0.11350898
2	control	work	T2	0.59592531
3	treatment	work	T2	0.35804998

arrange()

- sort by column values -ascending order by default
- arrange(column)
- arrange(desc(column)

arrange(color)

color	value
4	1
1	2
5	3
3	4
2	5



color	value
1	2
2	5
3	4
4	1
5	3

(Hadley Wickham)

mutate()

- add new columns from existing data

Formula: mutate(new_column = columnA - columnB)
mutate(new_column = columnA * columnB)
mutate(new_column = log2(columnA) / columnB)

mutate(minutes = time * 60)

id	trt	location	when	time
1	treatment	work	T1	0.08513597
2	control	work	T1	0.22543662
3	treatment	work	T1	0.27453052
4	control	work	T1	0.27230507
1	treatment	home	T1	0.61582931
2	control	home	T1	0.42967153
3	treatment	home	T1	0.65165567
4	control	home	T1	0.56773775
1	treatment	work	T2	0.11350898
2	control	work	T2	0.59592531
3	treatment	work	T2	0.35804998

id	trt	location	when	time	minutes
1	treatment	work	T1	0.08513597	5.1081582
2	control	work	T1	0.22543662	13.5261972
3	treatment	work	T1	0.27453052	16.4718312
4	control	work	T1	0.27230507	16.3383042
1	treatment	home	T1	0.61582931	36.9497586
2	control	home	T1	0.42967153	25.7802918
3	treatment	home	T1	0.65165567	39.0993402
4	control	home	T1	0.56773775	34.064265
1	treatment	work	T2	0.11350898	6.8105388
2	control	work	T2	0.59592531	35.7555186
3	treatment	work	T2	0.35804998	21.4829988

Try to:

- Compute speed in mph from (time) and distance (miles)
  - Which flight flew fastest?
- What was the longest flight delay in JFK in November?
- Which flights departed from LGA arrived to DTW early?

summarise()

- group_by() - ‘lock-in’ by variables
- aggregate/condense data


```
data %>%
    group_by(Organism, Treatment, Experiment) %>%
    summarise(      N = length(Count),
                 mean = mean(Count),
                   sd = sd(Count),
                   se = sd/sqrt(N))
```

Treatment	Experiment	Organism	Count
Antibiotic	1	Ecoli	285
Antibiotic	1	Ecoli	345
Antibiotic	1	Ecoli	298
Antibiotic	1	Ecoli	286
Antibiotic	1	Ecoli	354
None	1	Ecoli	146
None	1	Ecoli	180
None	1	Ecoli	137
None	1	Ecoli	179
None	1	Ecoli	168

Organism	Treatment	Experiment	N	mean	sd	se
Ecoli	Antibiotic	1	5	313.6	33.32116445	14.90167776
Ecoli	Antibiotic	2	5	351.6	36.66469692	16.39695094
Ecoli	Antibiotic	3	5	346.2	44.80736547	20.03846301
Ecoli	None	1	5	162	19.55760722	8.746427842
Ecoli	None	2	5	208.2	35.42880184	15.84424186
Ecoli	None	3	5	177.6	40.14722905	17.95438665

Try to determine:

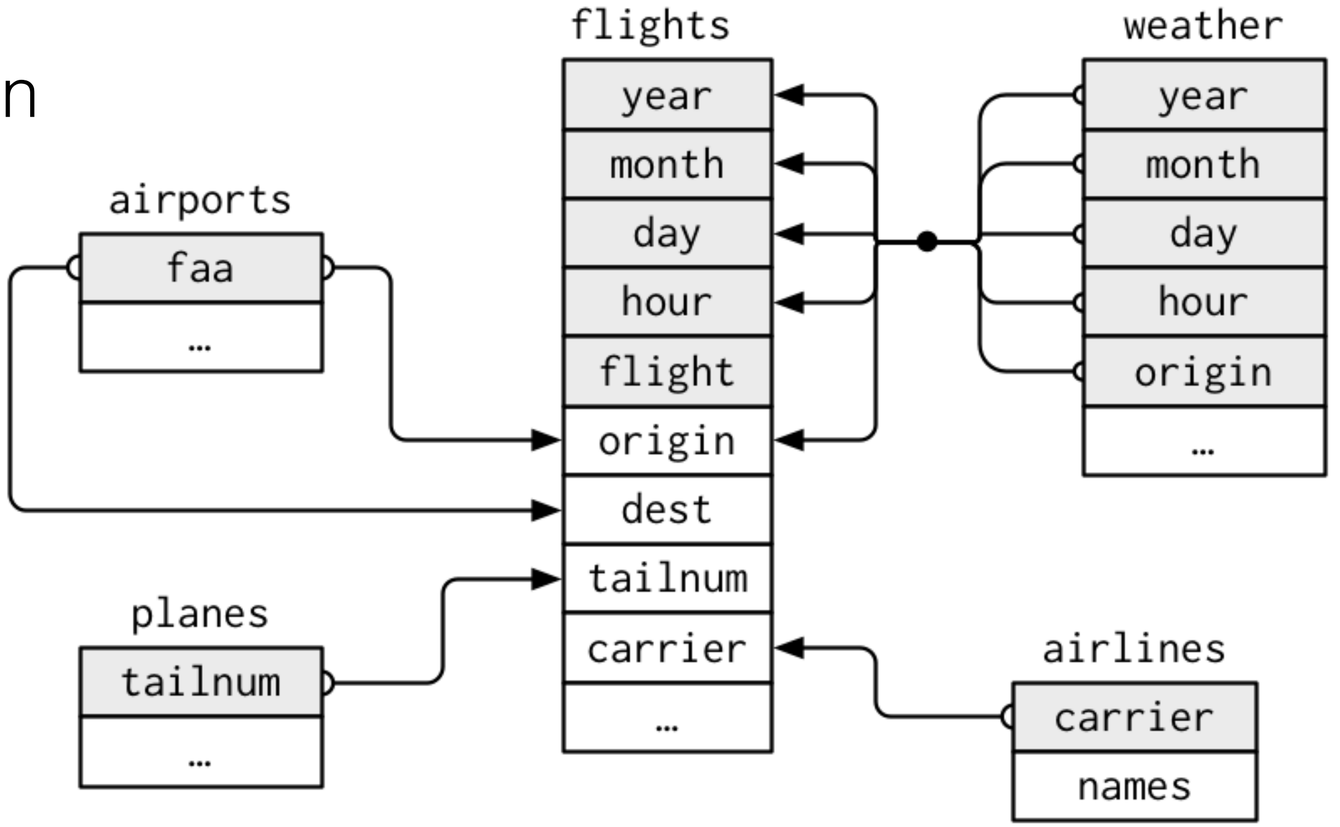
- Which airport had the most flights in December?
- Which NYC airport has the most airlines?
- How many United Airlines flights depart from JFK to ORD?

Demo!

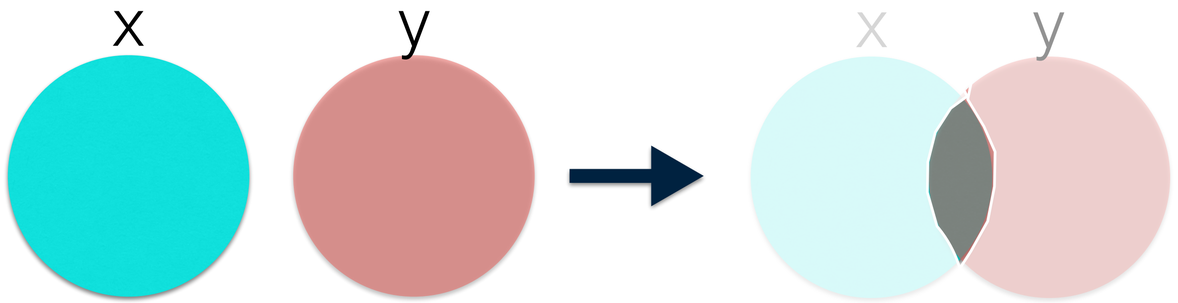
Joins

- Multiple datasets
- Merge relavent information
- Gain new information
- Blinded data

nycflights13 package



inner_join(x, y)



- combine things in common between x and y

superheroes

name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

publishers

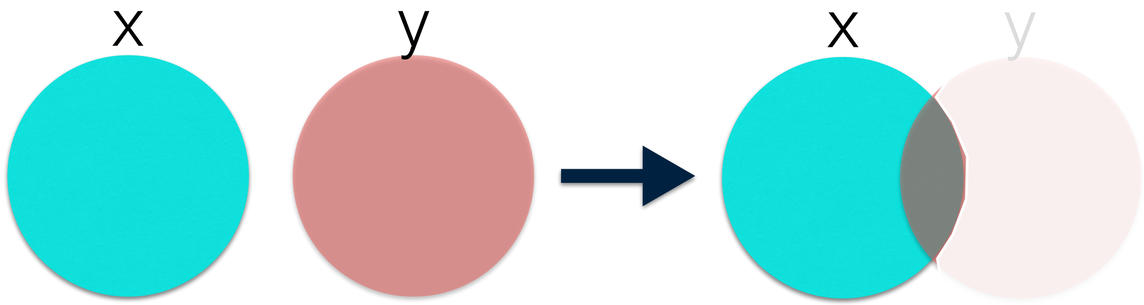
publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

inner_join(x = superheroes, y = publishers)

name	alignment	gender	publisher	yr_founded
Magneto	bad	male	Marvel	1939
Storm	good	female	Marvel	1939
Mystique	bad	female	Marvel	1939
Batman	good	male	DC	1934
Joker	bad	male	DC	1934
Catwoman	bad	female	DC	1934

(source:Jenny Bryan - Stat545)

left_join(x, y)



- Return all rows of x and all columns from x and y

superheroes

name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

publishers

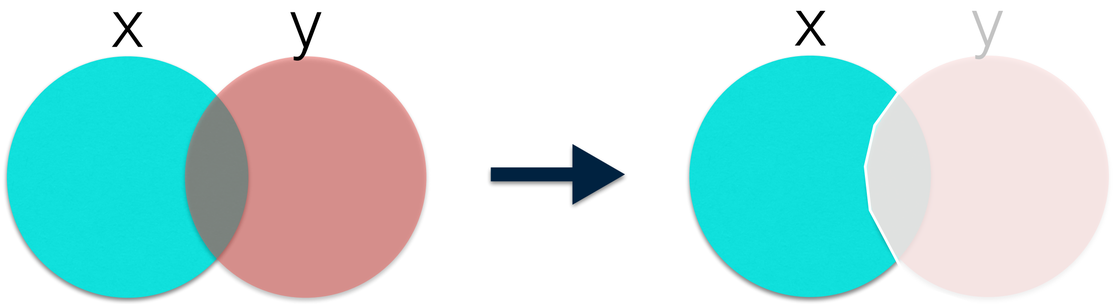
publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

left_join(x = superheroes, y = publishers)

name	alignment	gender	publisher	yr_founded
Magneto	bad	male	Marvel	1939
Storm	good	female	Marvel	1939
Mystique	bad	female	Marvel	1939
Batman	good	male	DC	1934
Joker	bad	male	DC	1934
Catwoman	bad	female	DC	1934
Hellboy	good	male	Dark Horse Comics	NA

(source:Jenny Bryan - Stat545)

anti_join(x, y)



- Keep what is distinct in x only

superheroes

name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

publishers

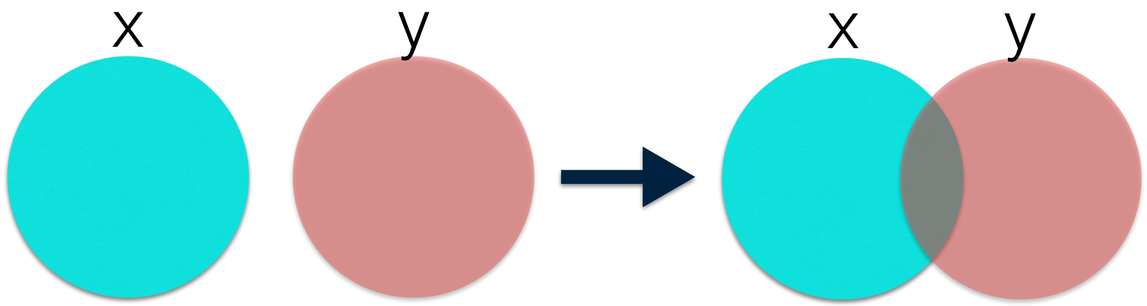
publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

anti_join(x = superheroes, y = publishers)

name	alignment	gender	publisher
Hellboy	good	male	Dark Horse Comics

(source:Jenny Bryan - Stat545)

full_join()



- Combine x and y, will introduce NAs

superheroes

name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

publishers

publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

full_join(x = superheroes, y = publishers)

name	alignment	gender	publisher	yr_founded
Magneto	bad	male	Marvel	1939
Storm	good	female	Marvel	1939
Mystique	bad	female	Marvel	1939
Batman	good	male	DC	1934
Joker	bad	male	DC	1934
Catwoman	bad	female	DC	1934
Hellboy	good	male	Dark Horse Comics	NA
NA	NA	NA	Image	1992

(source:Jenny Bryan - Stat545)

Demo!