Please keep rows
3 and 6 free again!

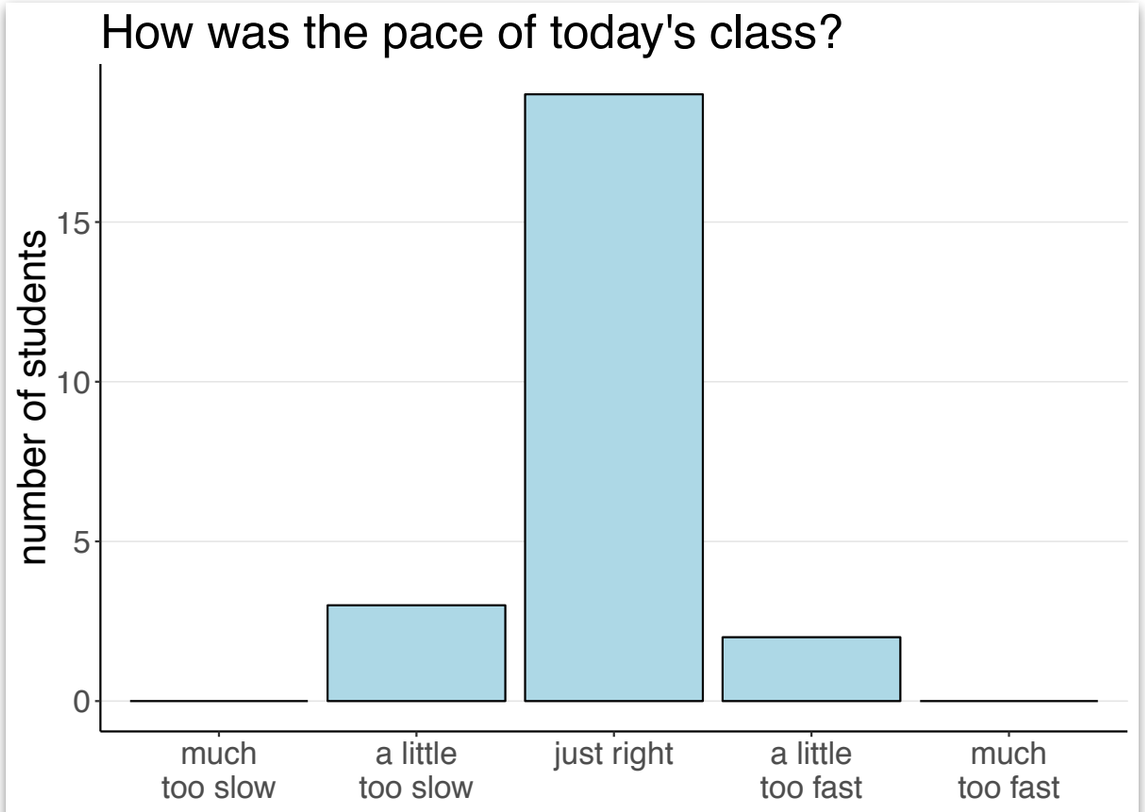
Data wrangling 2

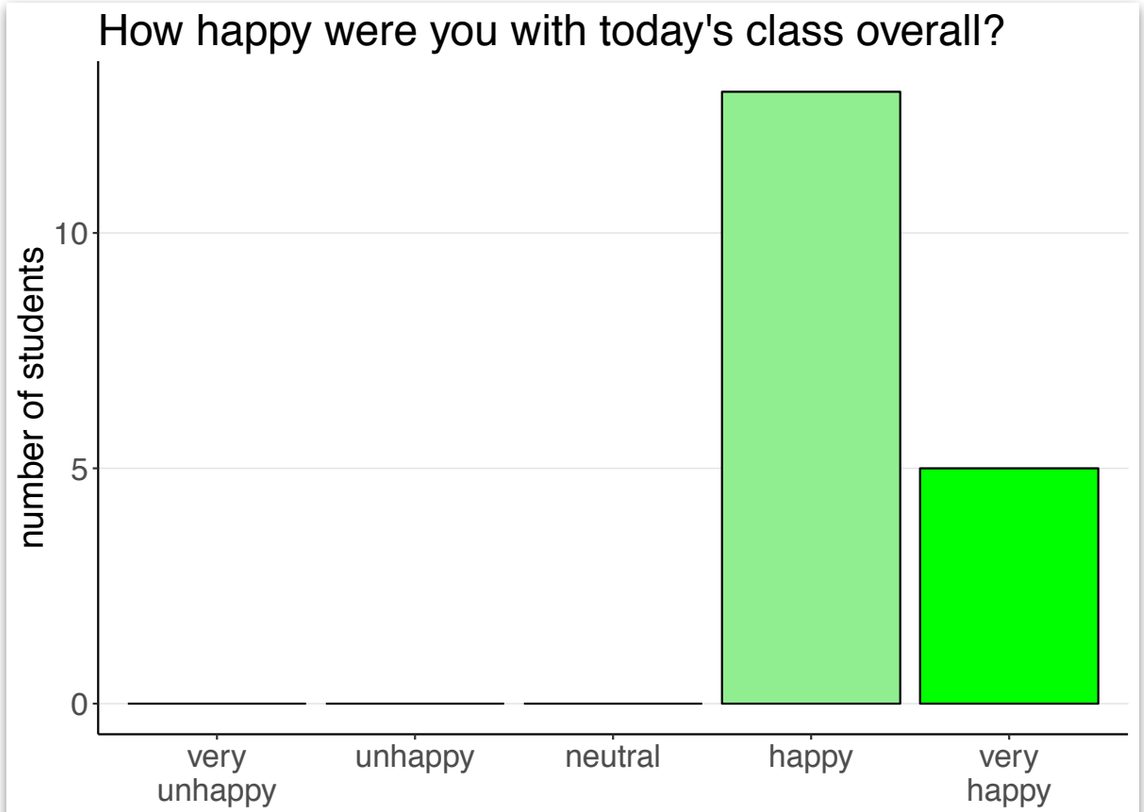
"Data don't make any sense,
we will have to resort to statistics."

01/16/2019

Your feedback

Your feedback

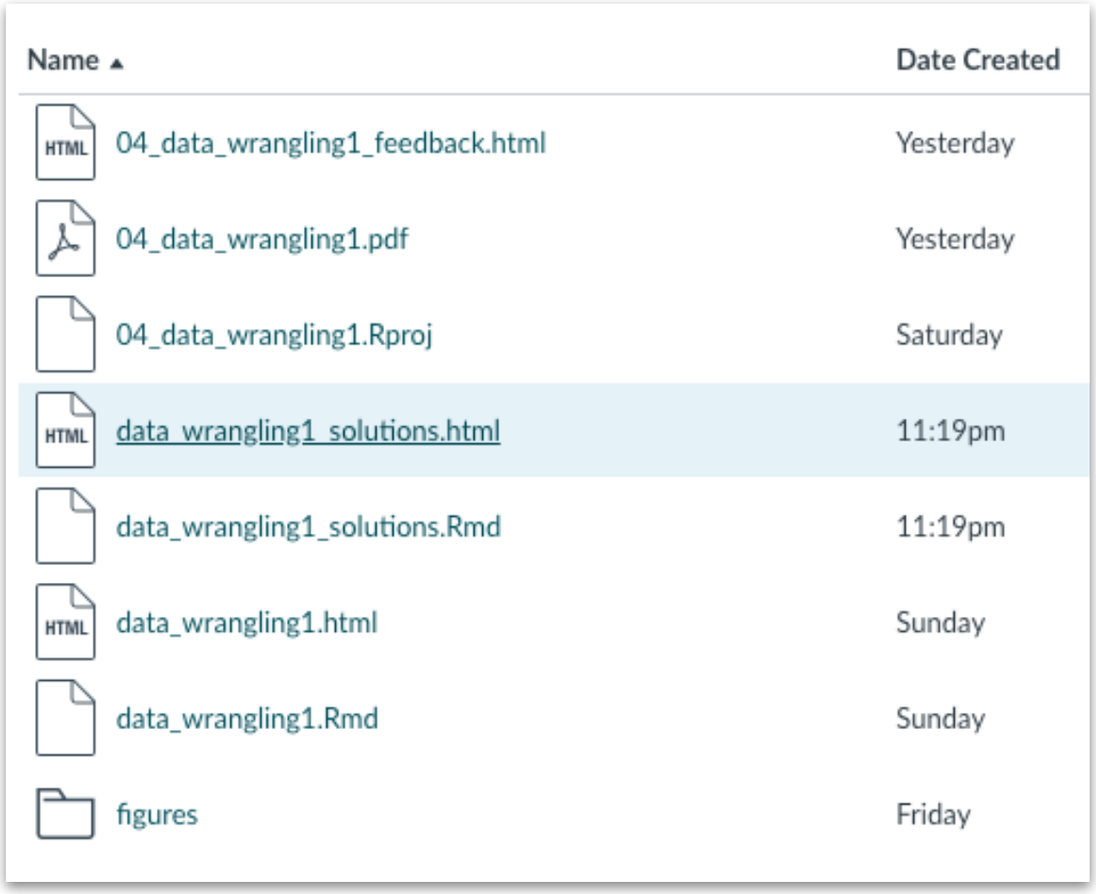




Your feedback

Everything is good and pace is just right. The only thing that I am concern is getting the practice answers from the lectures and sessions. Will it be possible for the lecturers and tutors **to post the practice answers online**, so that people who may have accidentally missed the answers can check it after classes? Thanks!

they are online now



1.2 Practice 1

Find out what the average `height` and `mass` (as well as the standard deviation) is from different `species` in different `homeworld`s. Why is the standard deviation `NA` for many groups?

```
df.starwars %>%
  group_by(species, homeworld) %>%
  summarize(mean_height = mean(height, na.rm = T),
            mean_mass = mean(mass, na.rm = T),
            sd_height = sd(height, na.rm = T),
            sd_mass = sd(mass, na.rm = T),
            n = n()) %>%
  ungroup()
```

```
# A tibble: 58 x 7
   species  homeworld   mean_height mean_mass sd_height sd_mass     n
   <chr>    <chr>             <dbl>     <dbl>     <dbl>   <dbl> <int>
 1 Aleena   Aleen Minor          79      15        NA      NA       1
 2 Besalisk Ojom                198     102        NA      NA       1
 3 Cerean   Cerea               198      82        NA      NA       1
 4 Chagrian Champala            196     NaN        NA      NA       1
 5 Clawdite Zolan               168      55        NA      NA       1
 6 Droid    Naboo                96      32        NA      NA       1
 7 Droid    Tatooine            132      53.5      49.5    30.4     2
 8 Droid    <NA>                200     140        NA      NA       2
 9 Dug      Malastare           112      40        NA      NA       1
10 Ewok     Endor                88      20        NA      NA       1
# … with 48 more rows
```

Logistics

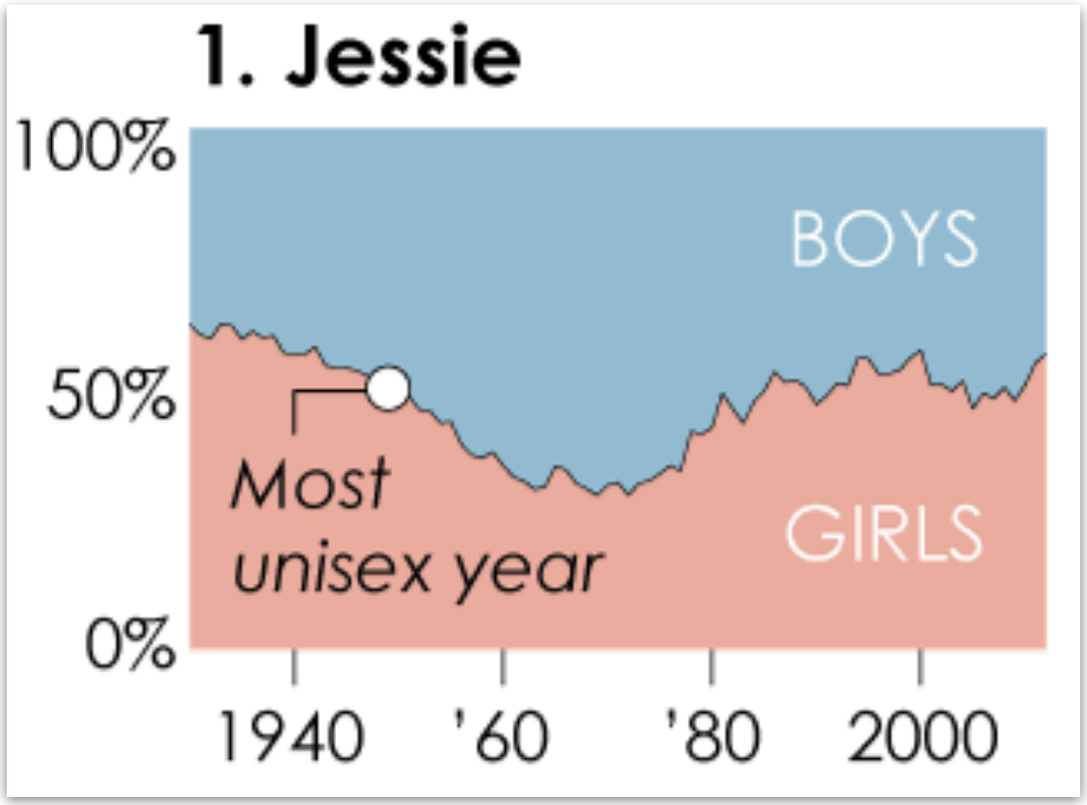
Homework 1

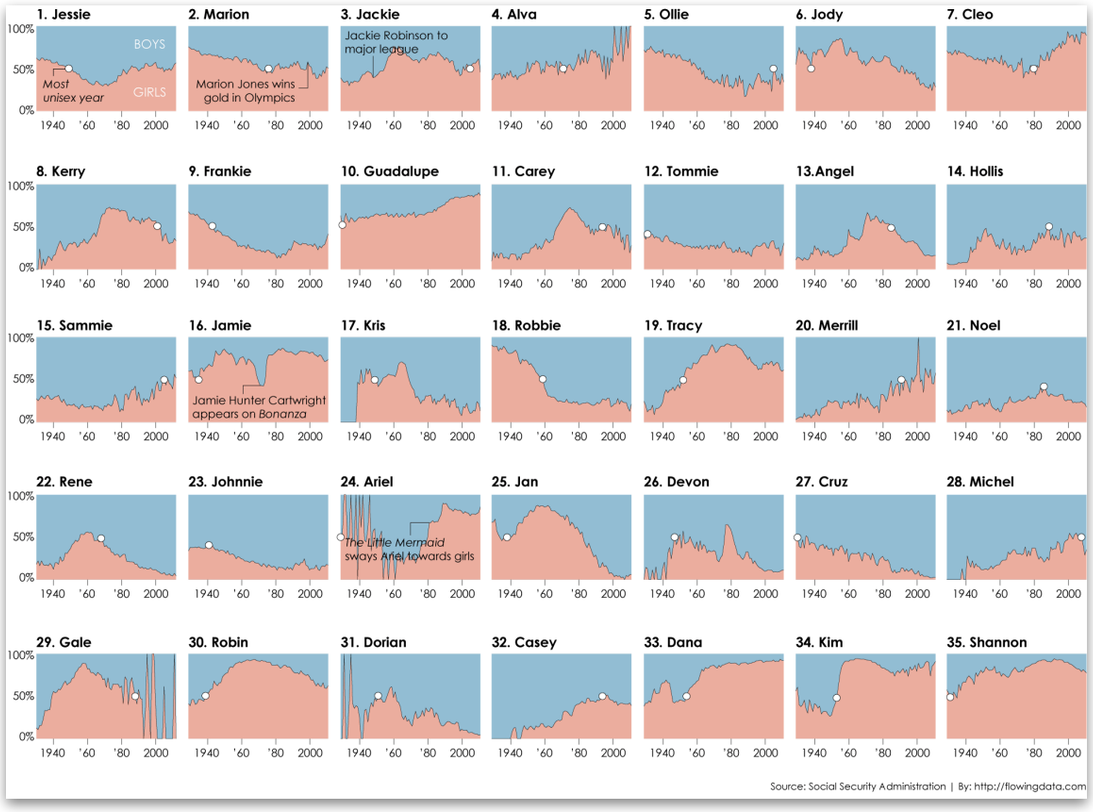
- **Make sure to submit on time!**
- From now onwards homework that is submitted late will count as **0 points**.
- Remember that your lowest-scored homework doesn't count.
- **Also**: You can submit early versions and then update later if you like (as long as it's before the deadline).
- Having homework submitted late makes our life difficult.
- It's good practice to take deadlines seriously!

Homework 2: Data wrangling and visualization

year	sex	name	n	prop
1880	F	Mary	7065	0.0723843285
1880	F	Anna	2604	0.0266792345
1880	F	Emma	2003	0.0205216999
1880	F	Elizabeth	1939	0.0198659891
1880	F	Minnie	1746	0.0178886111
1880	F	Margaret	1578	0.0161673702
1880	F	Ida	1472	0.0150813491
1880	F	Alice	1414	0.0144871112
1880	F	Bertha	1320	0.0135240359
1880	F	Sarah	1288	0.0131961805
1880	F	Annie	1258	0.0128888160
1880	F	Clara	1226	0.0125609606
1880	F	Ella	1156	0.0118437769
1880	F	Florence	1063	0.0108909471
1880	F	Cora	1045	0.0107065284
1880	F	Martha	1040	0.0106553010

- `babynames` data set from the "The United States Social Security Administration"
- **Copy the Master**



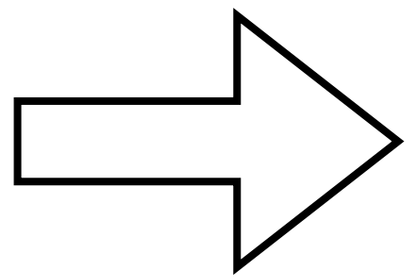


Data wrangling time ...

Replace missing values with mean

```
1 df.data %>%
2   group_by(participant) %>%
3   mutate(judgments = ifelse(test = is.na(judgments),
4                             yes = mean(judgments, na.rm = T),
5                             no = judgments))
```

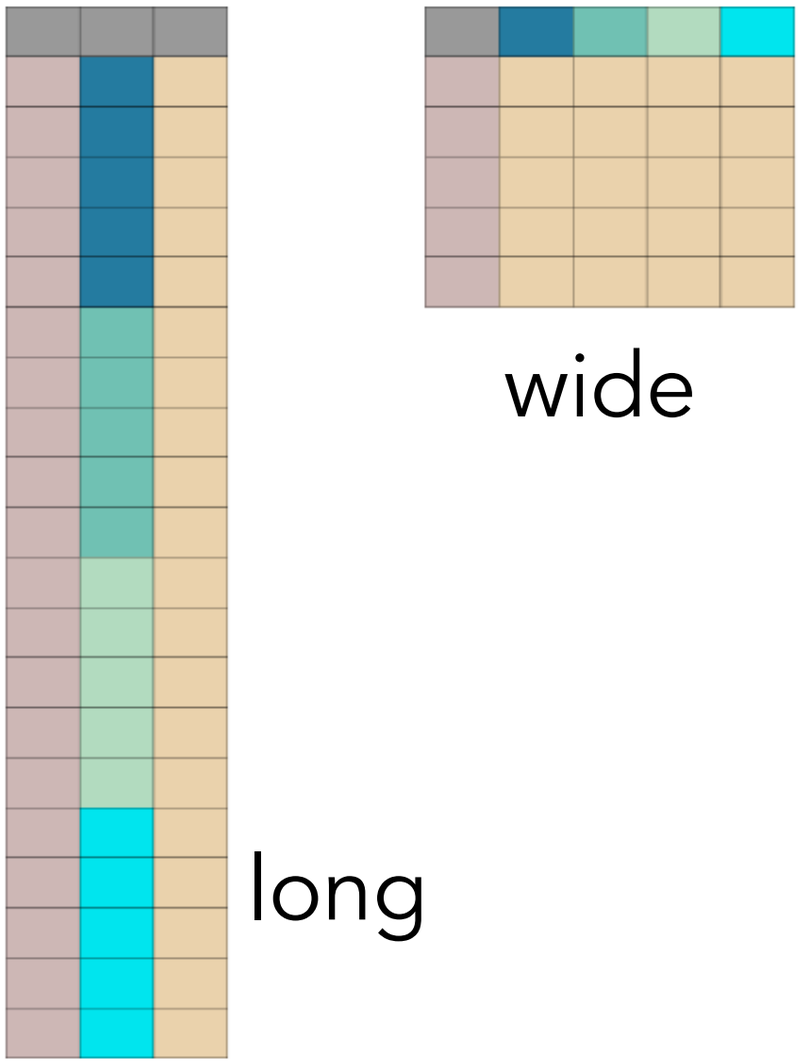
participant	judgments
1	66
1	NA
1	26
1	97
1	61
2	20
2	12
2	NA
2	85
2	55



participant	judgments
1	66.0
1	62.5
1	26.0
1	97.0
1	61.0
2	20.0
2	12.0
2	43.0
2	85.0
2	55.0

gather()

a b

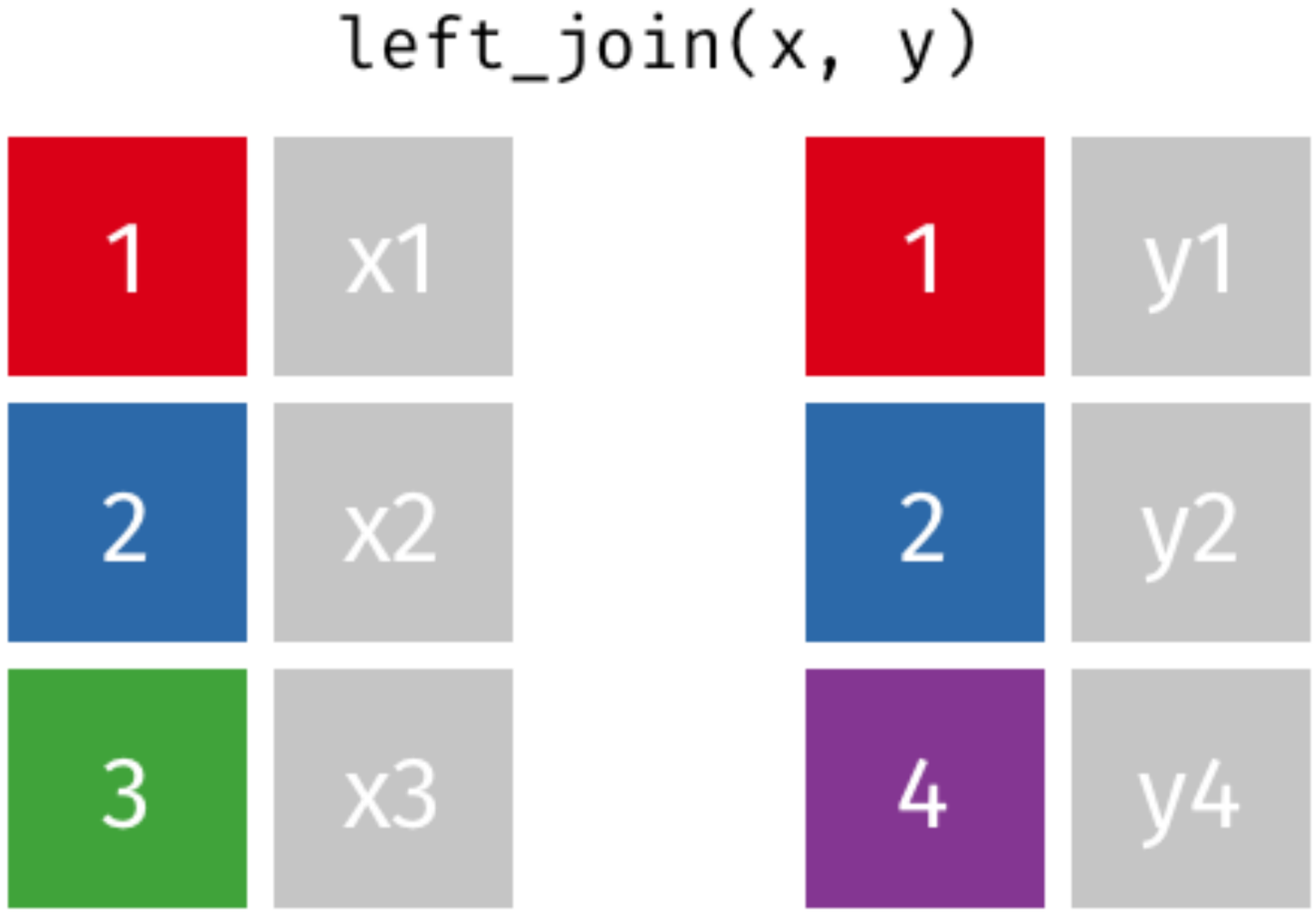


long

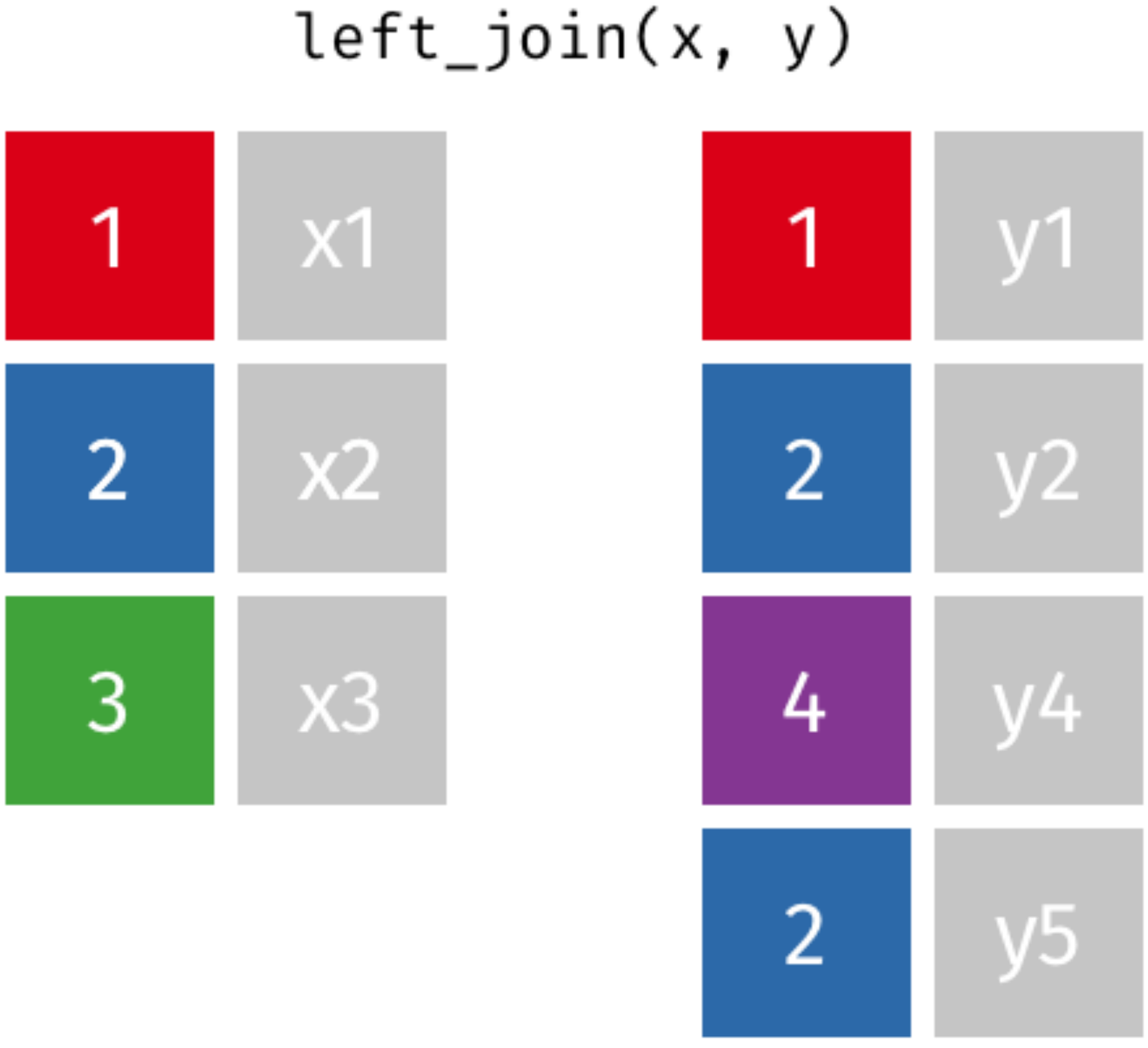
wide

spread()

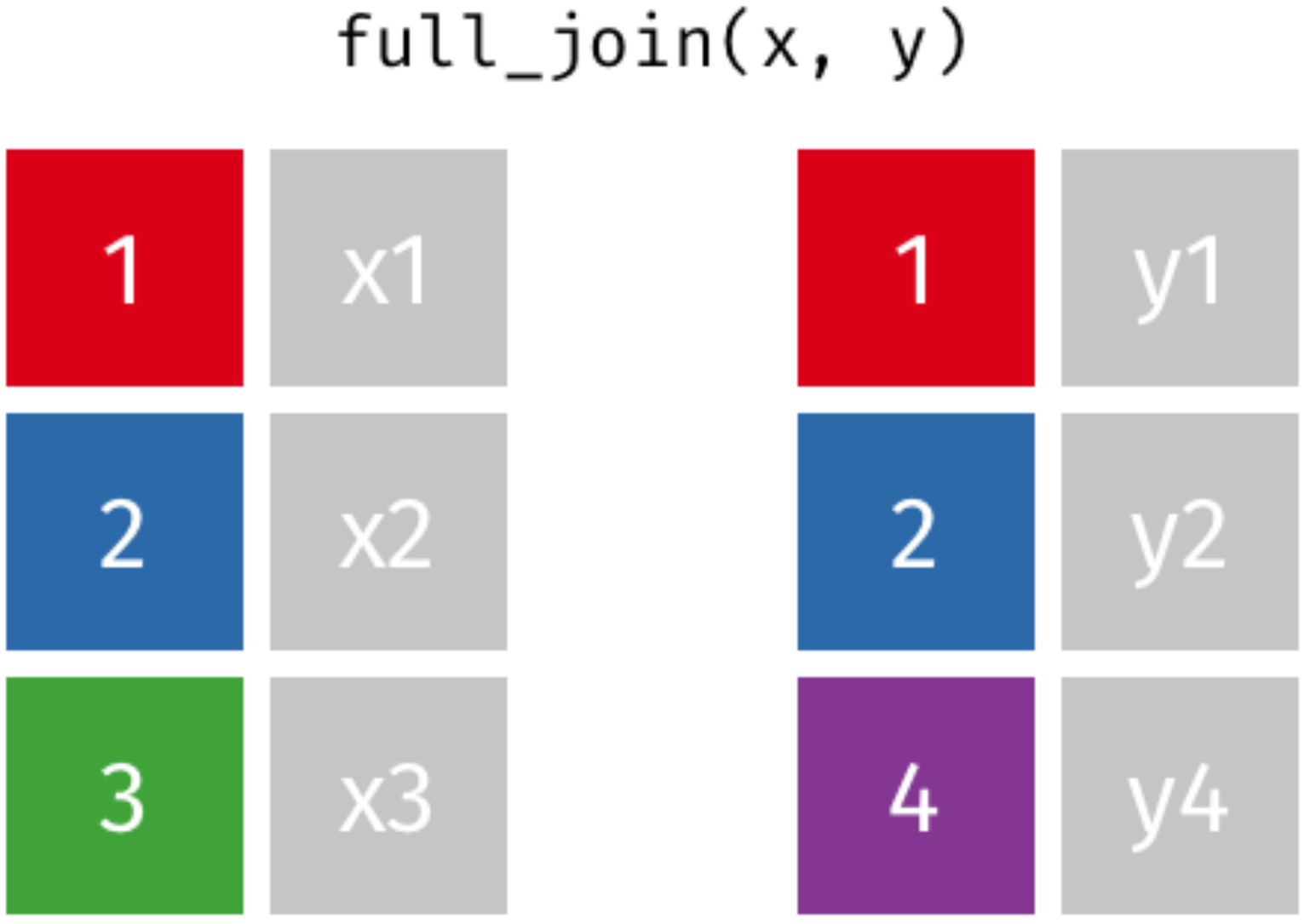
left_join()



left_join()



full_join()



Exercises

05:00

blue

pink

Thank you!