7: Combining Data Sets with dplyr

Suppose you have the following two data sets. The first, df1 has the variables id_numb and xvar. The second, df2 has the variables id and yvar. id_numb and id serve as identification variables, possibly with duplicates, where observations from the first data set with id_numb = 1 correspond to observations in the second data set with id = 1.

id_numb	xvar
1	16
1	-1
2	11
4	13

id	yvar
1	-1
2	-4
2	0
3	-9

Mutating Joins

- left_join()

dropped

left_join(df1, df2, by = c("id_numb" = "id"))

key in df1 key in df2

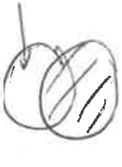
=>

id_numb	xvar	yvar
1	16	-1
1	-1	-1
2	11	-4
2	11	0
4	13	~~-9~~ NA

- right_join() (id=3 dropped)

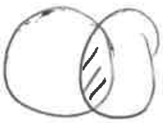
dropped right_join(df1, df2)

<=> left_join(df2, df1)



- inner_join()

inner_join(df1, df2, by=c("id_numb"="id"))

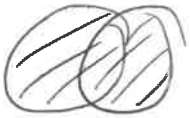


=>

id_numb	xvar
1	16
1	-1
2	11
4	13

id	yvar
1	-1
2	-4
2	0
3	-9

- full_join()



same as left_join() but adds

id_numb	xvar	yvar
3	NA	-9

Filtering Joins

- semi_join() same syntax

  dfl, df2 =>

keep anything that has a match in df2

id_numb	xvar
1	16
1	-1
2	11

- anti_join() same syntax

  dfl, df2 =>

id_numb	xvar
4	13

keep anything that doesn't have a match in df2.