MUNI SCI

M9DM2 Data mining 2

Databázové operace s datovými tabulkami v *R*

Ondřej Pokora

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita

8. 10. 2019

Relační data

- pomocí veličin v datových tabulkách popisují vztahy mezi pozorováními v tabulkách
- databázový jazyk SQL *(Structured Query Language)*
- knihovna `dplyr` přidává příkazy pro databázové operace nad datovými tabulkami v *R*
- knihovna `sqldf` přidává příkaz pro zápis SQL příkazů nad datovými tabulkami v *R*:

```
sqldf("SQL_prikaz")
```

Relační data

Proměnné (sloupce) v datových tabulkách se speciálním relačním významem:

- ***primary key*** – proměnná ve vlastní tabulce, unikátní identifikátor pozorování (řádku)
- ***foreign key*** – proměnná ve vlastní tabulce, unikátní identifikátor pozorování v jiné tabulce (tabulkách)
- jedna proměnná může být zároveň *primary key* i *foreign key*
- ***surrogate key*** – uměle doplněný primární klíč

Joins

left_join()

right_join()

inner_join()

full_join()

semi_join()

anti_join()

Binds

`bind_rows()`

`bind_cols()`

Spojování tabulek pomocí knihovny `dplyr`

Mutating joins – doplňují proměnné z `y` na konec `x`:

- `inner_join(x, y, by = "_")`
- `left_join(x, y, by = "_")`, `right_join (x, y, by = "_")`
- `full_join(x, y, by = "_")`

Filtering joins – vybírají pozorování z `x`:

- `semi_join(x, y, by = "_")`
- `anti_join(x, y, by = "_")`

Množinové operace:

- `union(x, y)`
- `intersect(x, y)`
- `setdiff(x, y)`

Svazování tabulek:

- `bind_rows(x, y)`, `rbind(x, y)`
- `bind_cols(x, y)`, `cbind(x, y)`

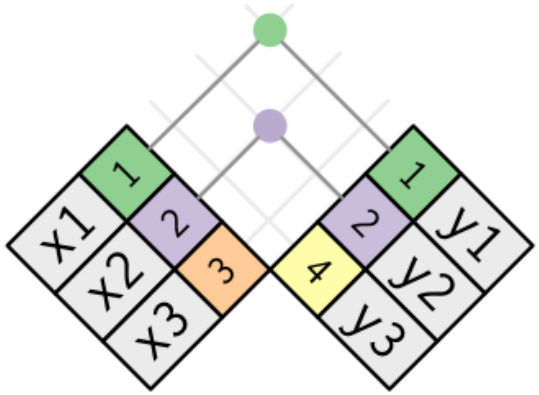
Inner join

x

1	x1
2	x2
3	x3

y

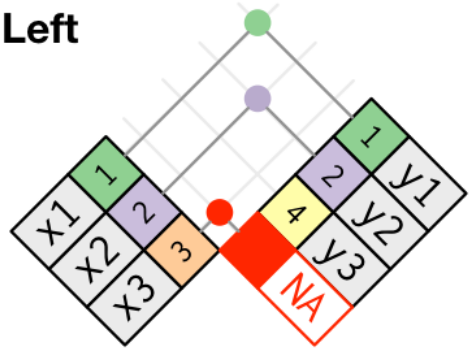
1	y1
2	y2
4	y3



key	val_x	val_y
1	x1	y1
2	x2	y2

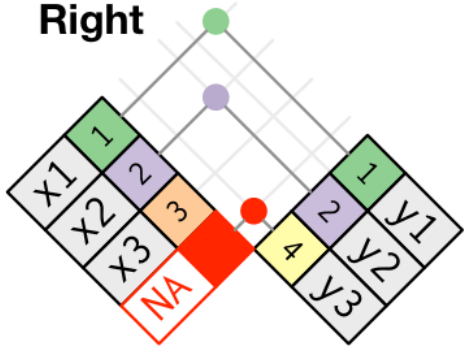
Outer joins

Left



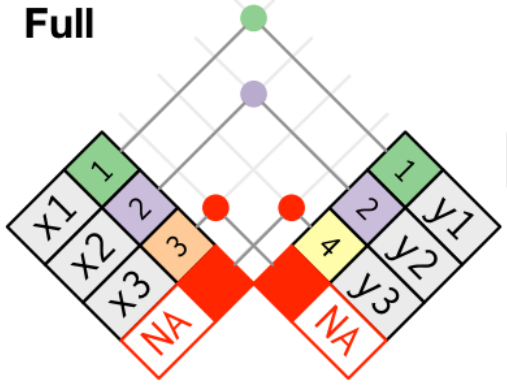
key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA

Right



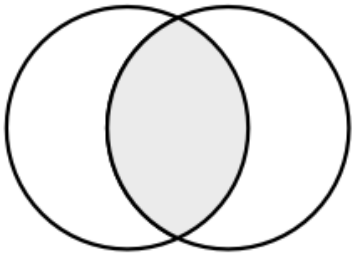
key	val_x	val_y
1	x1	y1
2	x2	y2
4	NA	y3

Full

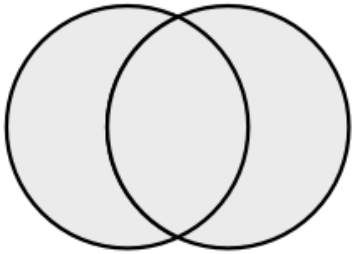


key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA
4	NA	y3

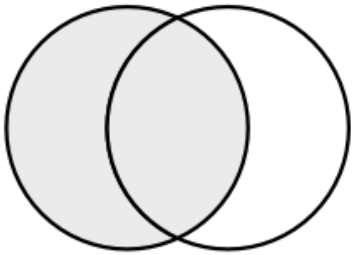
Inner join / Outer joins



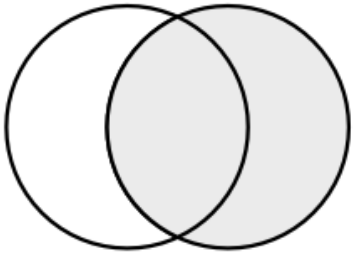
`inner_join(x, y)`



`full_join(x, y)`

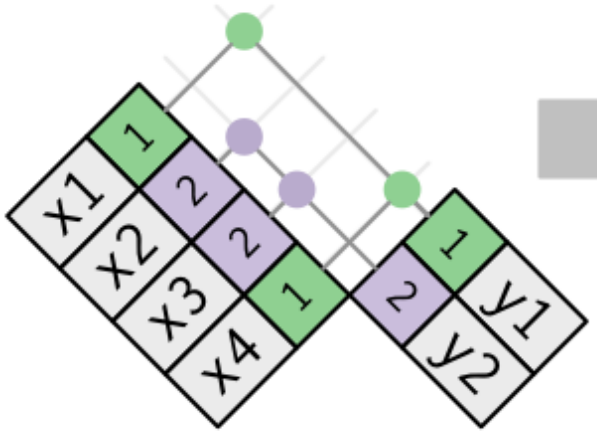


`left_join(x, y)`

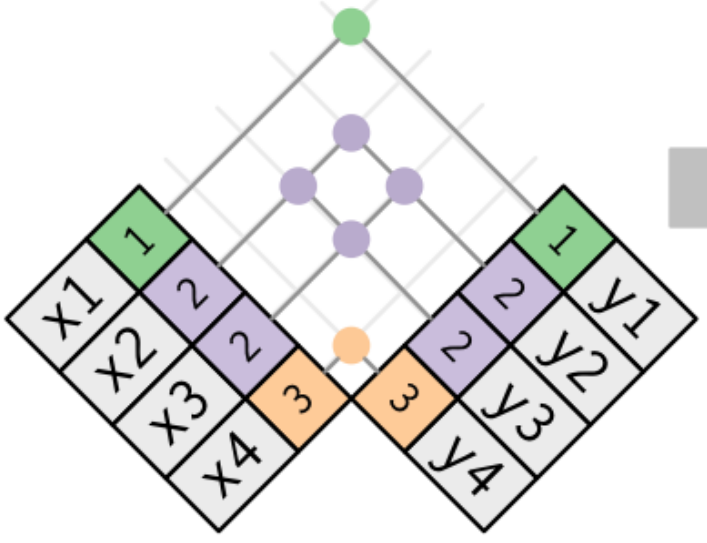


`right_join(x, y)`

Duplicitní hodnoty klíčů

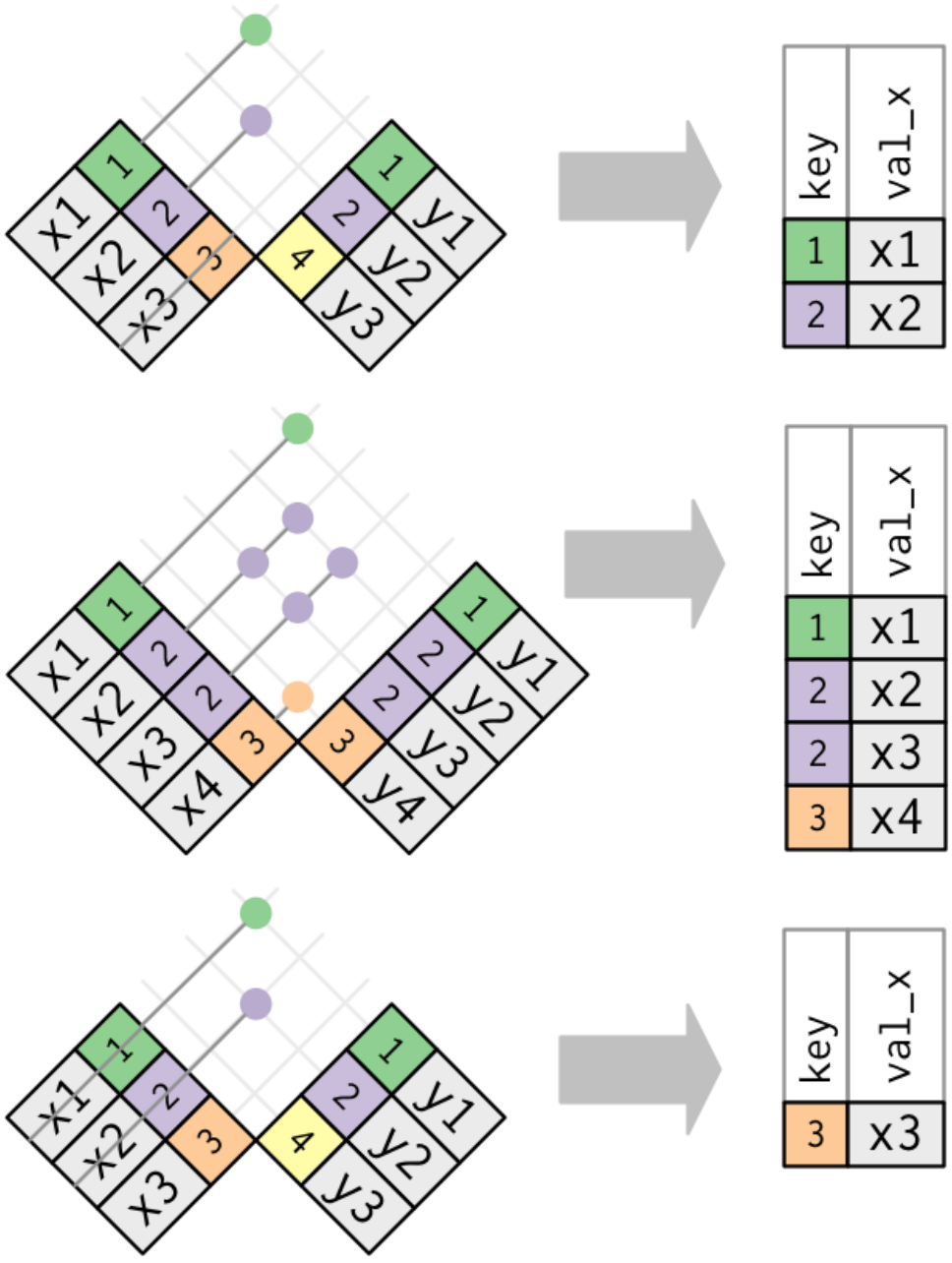


val_x	key	val_y
x1	1	y1
x2	2	y2
x3	2	y2
x4	1	y1

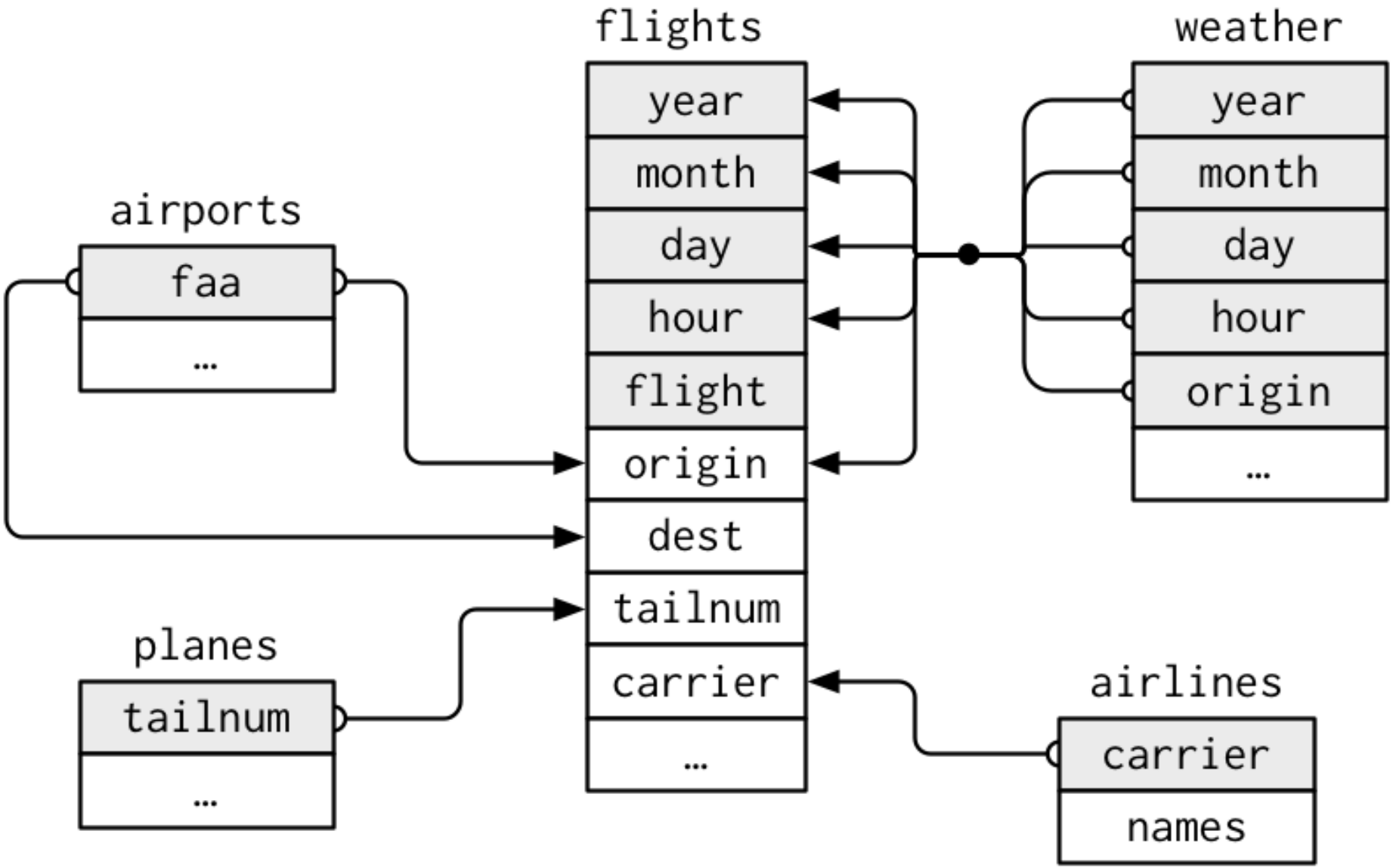


key	val_x	val_y
1	x1	y1
2	x2	y2
2	x2	y3
2	x3	y2
2	x3	y3
3	x4	y4

Semi join / Anti join



Příklad


flights
year
month
day
hour
flight
origin
dest
tailnum
carrier
…
weather
year
month
day
hour
origin
…
airports
faa
…
planes
tailnum
…
airlines
carrier
names