RUDN university

5100

Программирование в среде R

Шевцов Василий Викторович,
директор ДИТ РУДН, shevtsov_vv@rudn.university

Работа с реляционными данными

Данные

```
> flights
# A tibble: 336,776 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier flight tailnum origin
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>    <int> <chr>   <chr>
 1  2013     1     1      517            515         2      830            819        11 UA        1545 N14228  EWR
 2  2013     1     1      533            529         4      850            830        20 UA        1714 N24211  LGA
 3  2013     1     1      542            540         2      923            850        33 AA        1141 N619AA  JFK
 4  2013     1     1      544            545        -1     1004           1022       -18 B6         725 N804JB  JFK
 5  2013     1     1      554            600        -6      812            837       -25 DL         461 N668DN  LGA
 6  2013     1     1      554            558        -4      740            728        12 UA        1696 N39463  EWR
 7  2013     1     1      555            600        -5      913            854        19 B6         507 N516JB  EWR
 8  2013     1     1      557            600        -3      709            723       -14 EV        5708 N829AS  LGA
 9  2013     1     1      557            600        -3      838            846        -8 B6          79 N593JB  JFK
10  2013     1     1      558            600        -2      753            745         8 AA         301 N3ALAA  LGA
# ... with 336,766 more rows, and 6 more variables: dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
#   minute <dbl>, time_hour <dttm>
```

flights сведения о полетах

```
> airlines
# A tibble: 16 x 2
   carrier name
   <chr>   <chr>
 1 9E      Endeavor Air Inc.
 2 AA      American Airlines Inc.
 3 AS      Alaska Airlines Inc.
 4 B6      JetBlue Airways
 5 DL      Delta Air Lines Inc.
 6 EV      ExpressJet Airlines Inc.
 7 F9      Frontier Airlines Inc.
 8 FL      AirTran Airways Corporation
 9 HA      Hawaiian Airlines Inc.
10 MQ      Envoy Air
11 OO      SkyWest Airlines Inc.
12 UA      United Air Lines Inc.
13 US      US Airways Inc.
14 VX      Virgin America
15 WN      Southwest Airlines Co.
16 YV      Mesa Airlines Inc.
```

airlines справочник перевозчиков

5100

Данные

```
> airports
# A tibble: 1,458 x 8
   faa   name                              lat    lon   alt    tz dst   tzone
   <chr> <chr>                           <dbl>  <dbl> <dbl> <dbl> <chr> <chr>
 1 04G   Lansdowne Airport                41.1  -80.6  1044    -5 A     America/New_York
 2 06A   Moton Field Municipal Airport    32.5  -85.7   264    -6 A     America/Chicago
 3 06C   Schaumburg Regional
 4 06N   Randall Airport
 5 09J   Jekyll Island Airport
 6 0A9   Elizabethton Municipal Airport
 7 0G6   Williams County Airport
```

airports – информация о аэропортах
id – faa

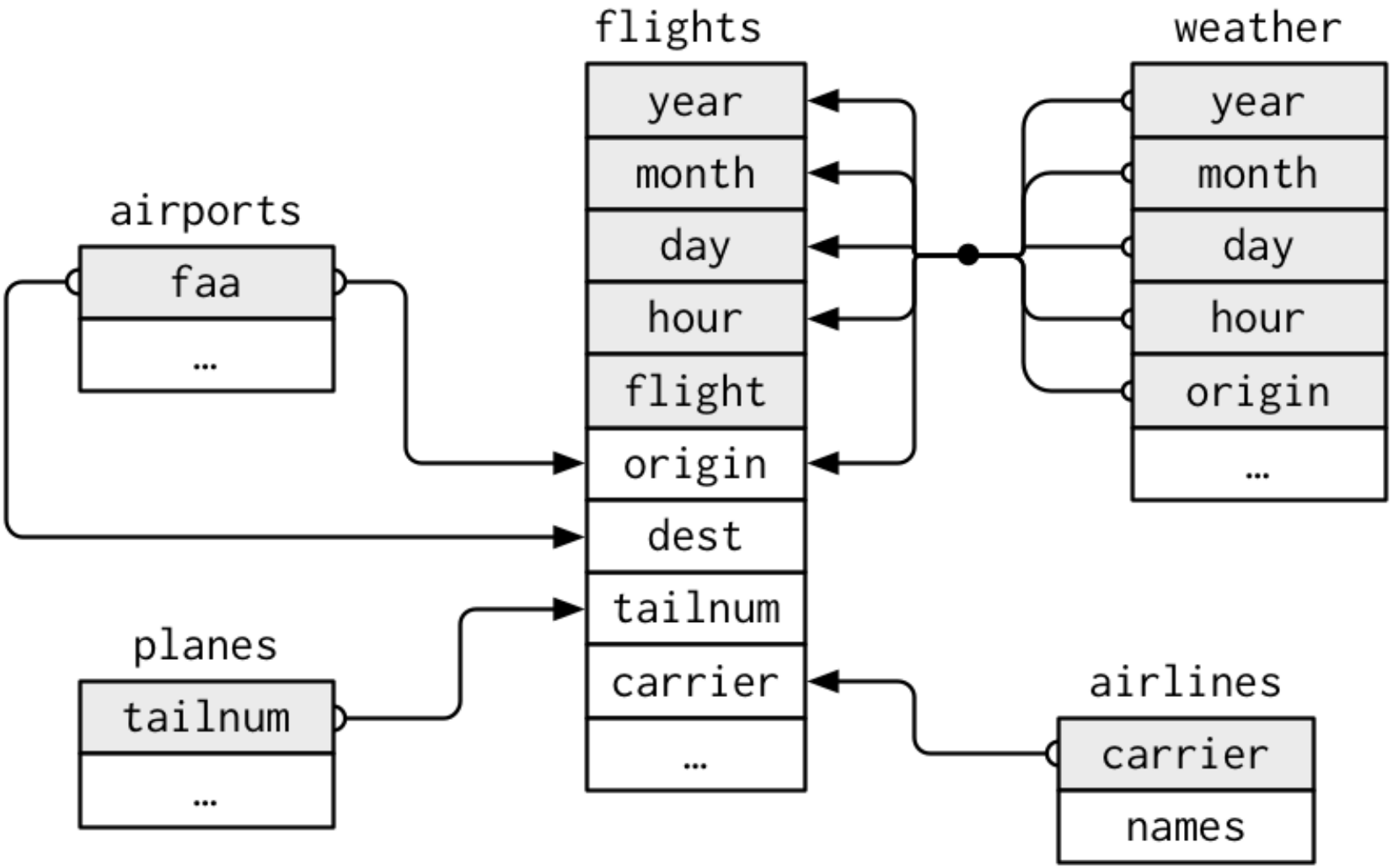
```
> planes
# A tibble: 3,322 x 9
   tailnum  year type                    manufacturer     model      engines seats speed engine
   <chr>   <int> <chr>                   <chr>            <chr>        <int> <int> <int> <chr>
 1 N10156   2004 Fixed wing multi engine EMBRAER          EMB-145XR        2    55    NA Turbo-fan
 2 N102UW   1998 Fixed wing multi engine AIRBUS INDUSTRIE A320-214        2   182    NA Turbo-fan
 3 N103US   1999 Fixed wing multi engine AIRBUS INDUSTRIE A320-214        2   182    NA Turbo-fan
 4 N104UW   1999 Fixed wing multi engine                                          fan
 5 N10575   2002 Fixed wing multi engine                                          fan
 6 N105UW   1999 Fixed wing multi engine                                          fan
 7 N107US   1999 Fixed wing multi engine                                          fan
 8 N108UW   1999 Fixed wing multi engine
```

planes – информация о самолетах
id – бортовой номер

```
> weather
# A tibble: 26,115 x 15
   origin  year month   day  hour  temp  dewp humid wind_dir wind_speed wind_gust precip pressure visib
   <chr>  <int> <int> <int> <int> <dbl> <dbl> <dbl>    <dbl>      <dbl>     <dbl>  <dbl>    <dbl> <dbl>
 1 EWR     2013     1     1     1  39.0  26.1  59.4      270       10.4        NA      0     1012    10
 2 EWR     2013     1     1     2  39.0  27.0  61.6      250        8.06       NA      0     1012.   10
 3 EWR     2013     1     1     3  39.0
 4 EWR     2013     1     1     4  39.9
 5 EWR     2013     1     1     5  39.0
 6 EWR     2013     1     1     6  37.9
 7 EWR     2013     1     1     7  39.0  28.0  64.4      240       15.0        NA      0     1012.   10
 8 EWR     2013     1     1     8  39.9  28.0  62.2      250       10.4        NA      0     1012.   10
 9 EWR     2013     1     1     9  39.9  28.0  62.2      260       15.0        NA      0     1013.   10
10 EWR     2013     1     1    10  41    28.0  59.6      260       13.8        NA      0     1012.   10
# ... with 26,105 more rows, and 1 more variable: time_hour <dttm>
```

weather – почасовые сведения о погодных условиях

Связи между таблицами



Связи между таблицами

- flights соединяется с planes посредством ключа tailnum
- flights соединяется с airlines посредством ключа carrier
- flights соединяется с airports посредством двух ключей origin, dest
- flights соединяется с weather посредством ключей origin, year, month, day, hour.

- Первичный ключ
- Внешний ключ

Операции с ключами

```
planes %>%
  count(tailnum) %>%
  filter(n > 1)
```

```
# A tibble: 0 x 2
# ... with 2 variables: tailnum <chr>, n <int>
```

```
weather %>%
  count(year, month, day, hour, origin) %>%
  filter(n > 1)
```

```
# A tibble: 3 x 6
   year month   day  hour origin     n
  <int> <int> <int> <int> <chr>  <int>
1  2013    11     3     1 EWR        2
2  2013    11     3     1 JFK        2
3  2013    11     3     1 LGA        2
```

```
flights %>%
count(year, month, day, flight) %>%
filter(n > 1)
```

```
# A tibble: 29,768 x 5
    year month   day flight     n
   <int> <int> <int>  <int> <int>
 1  2013     1     1      1     2
 2  2013     1     1      3     2
 3  2013     1     1      4     2
 4  2013     1     1     11     3
 5  2013     1     1     15     2
 6  2013     1     1     21     2
 7  2013     1     1     27     4
 8  2013     1     1     31     2
 9  2013     1     1     32     2
10  2013     1     1     35     2
# ... with 29,758 more rows
```

```
flights %>%
count(year, month, day, tailnum) %>%
filter(n > 1)
```

```
# A tibble: 64,928 x 5
    year month   day tailnum     n
   <int> <int> <int> <chr>   <int>
 1  2013     1     1 N0EGMQ      2
 2  2013     1     1 N11189      2
 3  2013     1     1 N11536      2
 4  2013     1     1 N11544      3
 5  2013     1     1 N11551      2
 6  2013     1     1 N12540      2
 7  2013     1     1 N12567      2
 8  2013     1     1 N13123      2
 9  2013     1     1 N13538      3
10  2013     1     1 N13566      3
# ... with 64,918 more rows
```

RUDN university

5100

Связи

```
flights2 <- flights %>%
  select(year:day, hour, origin, dest, tailnum, carrier)
```

```
> flights2
# A tibble: 336,776 x 8
    year month   day  hour origin dest  tailnum carrier
   <int> <int> <int> <dbl> <chr>  <chr> <chr>   <chr>
 1  2013     1     1     5 EWR    IAH   N14228  UA
 2  2013     1     1     5 LGA    IAH   N24211  UA
 3  2013     1     1     5 JFK    MIA   N619AA  AA
 4  2013     1     1     5 JFK    BQN   N804JB  B6
 5  2013     1     1     6 LGA    ATL   N668DN  DL
 6  2013     1     1     5 EWR    ORD   N39463  UA
 7  2013     1     1     6 EWR    FLL   N516JB  B6
 8  2013     1     1     6 LGA    IAD   N829AS  EV
 9  2013     1     1     6 JFK    MCO   N593JB  B6
10  2013     1     1     6 LGA    ORD   N3ALAA  AA
# ... with 336,766 more rows
```

```
flights2 %>%
  select(-origin, -dest) %>%
  left_join(airlines, by = "carrier")
```

To join by different variables on x and y use a named vector. For example, by = c("a" = "b") will match x.a to y.b.

```
> flights2
# A tibble: 336,776 x 8
    year month   day  hour origin dest  tailnum carrier
   <int> <int> <int> <dbl> <chr>  <chr> <chr>   <chr>
 1  2013     1     1     5 EWR    IAH   N14228  UA
 2  2013     1     1     5 LGA    IAH   N24211  UA
 3  2013     1     1     5 JFK    MIA   N619AA  AA
 4  2013     1     1     5 JFK    BQN   N804JB  B6
 5  2013     1     1     6 LGA    ATL   N668DN  DL
 6  2013     1     1     5 EWR    ORD   N39463  UA
 7  2013     1     1     6 EWR    FLL   N516JB  B6
 8  2013     1     1     6 LGA    IAD   N829AS  EV
 9  2013     1     1     6 JFK    MCO   N593JB  B6
10  2013     1     1     6 LGA    ORD   N3ALAA  AA
# ... with 336,766 more rows
```

Связи

```
flights2 %>%
select(-origin, -dest) %>%
mutate(name = airlines$name[match(carrier, airlines$carrier)])
```

```
# A tibble: 336,776 x 7
    year month   day  hour tailnum carrier name
   <int> <int> <int> <dbl> <chr>   <chr>   <chr>
 1  2013     1     1     5 N14228  UA      United Air Lines Inc.
 2  2013     1     1     5 N24211  UA      United Air Lines Inc.
 3  2013     1     1     5 N619AA  AA      American Airlines Inc.
 4  2013     1     1     5 N804JB  B6      JetBlue Airways
 5  2013     1     1     6 N668DN  DL      Delta Air Lines Inc.
 6  2013     1     1     5 N39463  UA      United Air Lines Inc.
 7  2013     1     1     6 N516JB  B6      JetBlue Airways
 8  2013     1     1     6 N829AS  EV      ExpressJet Airlines Inc.
 9  2013     1     1     6 N593JB  B6      JetBlue Airways
10  2013     1     1     6 N3ALAA  AA      American Airlines Inc.
# ... with 336,766 more rows
```

match(x,table), x %in% table — выполняет поиск элементов в векторе table, которые совпадают со значениями из вектора x

Связи

Исходные данные: ключ, значение

```
x <- tribble(
  ~key, ~val_x,
  1, "x1",
  2, "x2",
  3, "x3"
)
```

```
> x
# A tibble: 3 x 2
    key val_x
  <dbl> <chr>
1     1 x1
2     2 x2
3     3 x3
```

x

1	x1
2	x2
3	x3

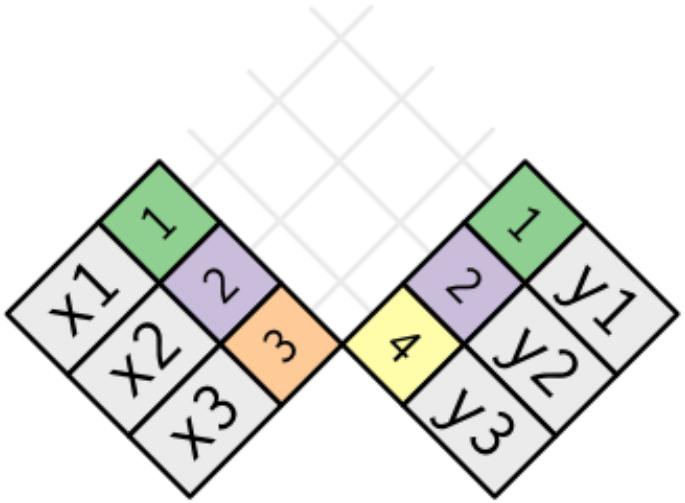
y

1	y1
2	y2
4	y3

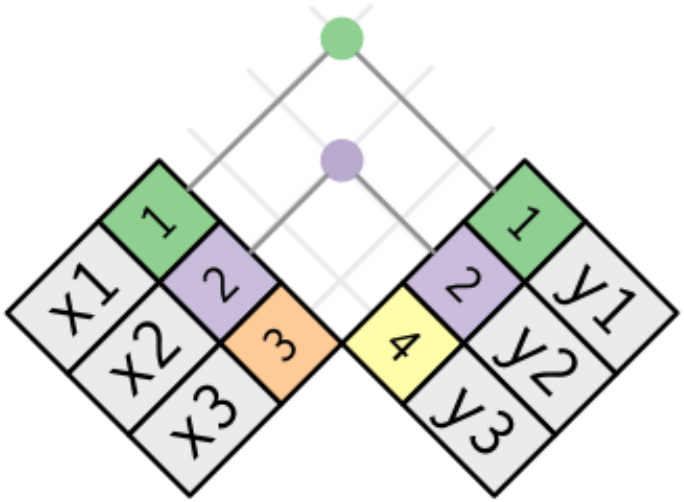
```
y <- tribble(
~key, ~val_y,
1, "y1",
2, "y2",
4, "y3"
)
```

```
> y
# A tibble: 3 x 2
    key val_y
  <dbl> <chr>
1     1 y1
2     2 y2
3     4 y3
```

Связи будут отображаться в виде пересечений пары линий



Inner Join (внутреннее соединение)



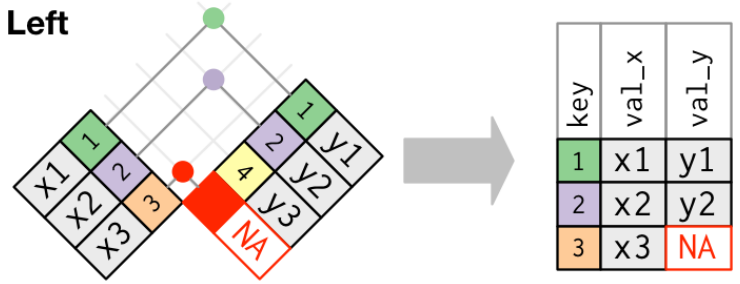
key	val_x	val_y
1	x1	y1
2	x2	y2

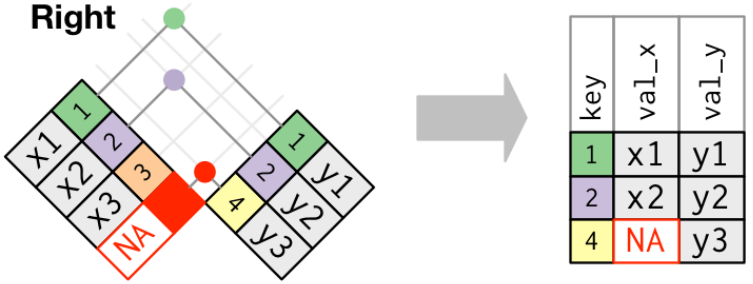
```
x %>%
  inner_join(y, by = "key")
```

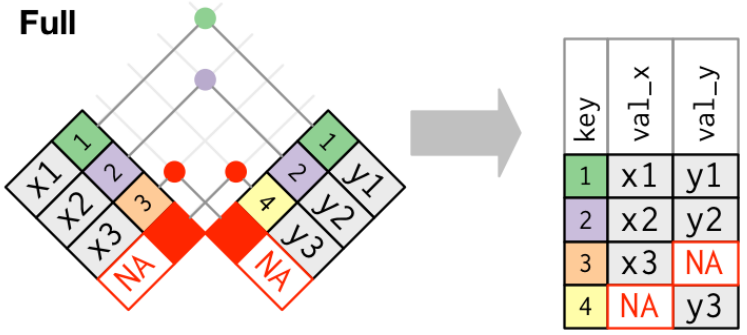
```
# A tibble: 2 x 3
    key val_x val_y
  <dbl> <chr> <chr>
1     1 x1    y1
2     2 x2    y2
```

Outer Joins (внешнее соединение)

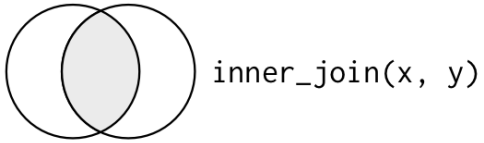
- **left join** сохраняет все значения в x
- **right join** сохраняет все значения в y
- **full join** сохраняет все значения в x и y

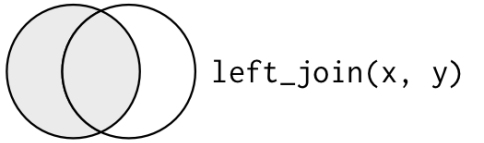


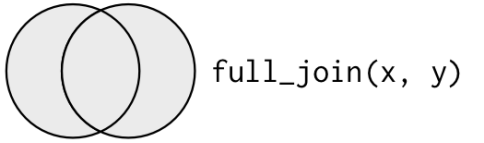


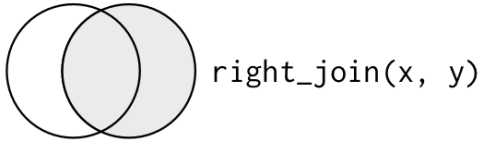


Диаграммы Венна



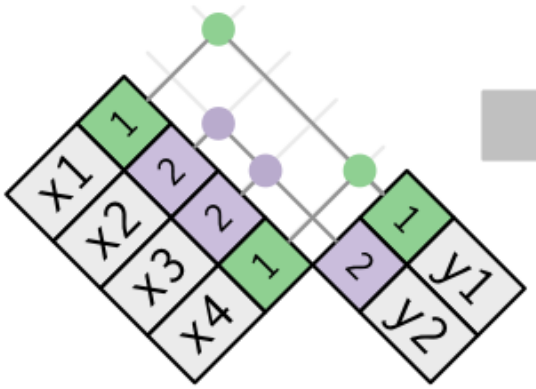






Неуникальные ключи

```
x <- tribble(
  ~key, ~val_x,
  1, "x1",
  2, "x2",
  2, "x3",
  1, "x4"
)
y <- tribble(
  ~key, ~val_y,
  1, "y1",
  2, "y2"
)
left_join(x, y, by = "key")
```

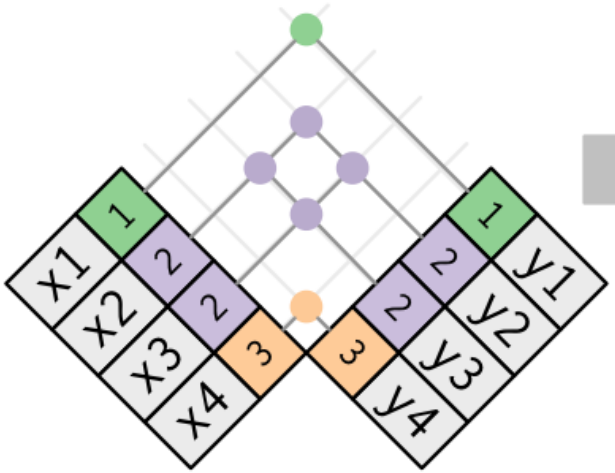


val_x	key	val_y
x1	1	y1
x2	2	y2
x3	2	y2
x4	1	y1

```
# A tibble: 4 x 3
    key val_x val_y
  <dbl> <chr> <chr>
1     1 x1    y1
2     2 x2    y2
3     2 x3    y2
4     1 x4    y1
```

Неуникальные ключи

```
x <- tribble(
  ~key, ~val_x,
  1, "x1",
  2, "x2",
  2, "x3",
  3, "x4"
)
y <- tribble(
  ~key, ~val_y,
  1, "y1",
  2, "y2",
  2, "y3",
  3, "y4"
)
left_join(x, y, by = "key")
```



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

```
# A tibble: 6 x 3
    key val_x val_y
  <dbl> <chr> <chr>
1     1 x1    y1
2     2 x2    y2
3     2 x2    y3
4     2 x3    y2
5     2 x3    y3
6     3 x4    y4
```

Определение ключевых столбцов. by = NULL

```
flights2 %>%
        left_join(weather)
```

```
Joining, by = c("year", "month", "day", "hour", "origin")
# A tibble: 336,776 x 18
    year month   day  hour origin dest  tailnum carrier  temp  dewp humid wind_dir wind_speed
   <int> <int> <int> <dbl> <chr>  <chr> <chr>   <chr>   <dbl> <dbl> <dbl>    <dbl>      <dbl>
 1  2013     1     1     5 EWR    IAH   N14228  UA       39.0  28.0  64.4      260       12.7
 2  2013     1     1     5 LGA    IAH   N24211  UA       39.9  25.0  54.8      250       15.0
 3  2013     1     1     5 JFK    MIA   N619AA  AA       39.0  27.0  61.6      260       15.0
 4  2013     1     1     5 JFK    BQN   N804JB  B6       39.0  27.0  61.6      260       15.0
 5  2013     1     1     6 LGA    ATL   N668DN  DL       39.9  25.0  54.8      260       16.1
 6  2013     1     1     5 EWR    ORD   N39463  UA       39.0  28.0  64.4      260       12.7
 7  2013     1     1     6 EWR    FLL   N516JB  B6       37.9  28.0  67.2      240       11.5
 8  2013     1     1     6 LGA    IAD   N829AS  EV       39.9  25.0  54.8      260       16.1
 9  2013     1     1     6 JFK    MCO   N593JB  B6       37.9  27.0  64.3      260       13.8
10  2013     1     1     6 LGA    ORD   N3ALAA  AA       39.9  25.0  54.8      260       16.1
# ... with 336,766 more rows, and 5 more variables: wind_gust <dbl>, precip <dbl>,
#   pressure <dbl>, visib <dbl>, time_hour <dttm>
```

Определение ключевых столбцов. by = "x"

```
flights2 %>%
        left_join(planes, by = "tailnum")
```

```
# A tibble: 336,776 x 16
   year.x month   day  hour origin dest  tailnum carrier year.y type  manufacturer model
    <int> <int> <int> <dbl> <chr>  <chr> <chr>   <chr>    <int> <chr> <chr>        <chr>
 1   2013     1     1     5 EWR    IAH   N14228  UA        1999 Fixe~ BOEING       737-~
 2   2013     1     1     5 LGA    IAH   N24211  UA        1998 Fixe~ BOEING       737-~
 3   2013     1     1     5 JFK    MIA   N619AA  AA        1990 Fixe~ BOEING       757-~
 4   2013     1     1     5 JFK    BQN   N804JB  B6        2012 Fixe~ AIRBUS       A320~
 5   2013     1     1     6 LGA    ATL   N668DN  DL        1991 Fixe~ BOEING       757-~
 6   2013     1     1     5 EWR    ORD   N39463  UA        2012 Fixe~ BOEING       737-~
 7   2013     1     1     6 EWR    FLL   N516JB  B6        2000 Fixe~ AIRBUS INDU~ A320~
 8   2013     1     1     6 LGA    IAD   N829AS  EV        1998 Fixe~ CANADAIR     CL-6~
 9   2013     1     1     6 JFK    MCO   N593JB  B6        2004 Fixe~ AIRBUS       A320~
10   2013     1     1     6 LGA    ORD   N3ALAA  AA          NA NA    NA           NA
# ... with 336,766 more rows, and 4 more variables: engines <int>, seats <int>, speed <int>,
#   engine <chr>
```

Определение ключевых столбцов. by = c("a"="b")

```
flights2 %>%
        left_join(airports, c("origin" = "faa"))
```

```
# A tibble: 336,776 x 15
    year month   day  hour origin dest  tailnum carrier name    lat   lon   alt    tz dst
   <int> <int> <int> <dbl> <chr>  <chr> <chr>   <chr>   <chr> <dbl> <dbl> <dbl> <dbl> <chr>
 1  2013     1     1     5 EWR    IAH   N14228  UA      Newa~  40.7 -74.2    18    -5 A
 2  2013     1     1     5 LGA    IAH   N24211  UA      La G~  40.8 -73.9    22    -5 A
 3  2013     1     1     5 JFK    MIA   N619AA  AA      John~  40.6 -73.8    13    -5 A
 4  2013     1     1     5 JFK    BQN   N804JB  B6      John~  40.6 -73.8    13    -5 A
 5  2013     1     1     6 LGA    ATL   N668DN  DL      La G~  40.8 -73.9    22    -5 A
 6  2013     1     1     5 EWR    ORD   N39463  UA      Newa~  40.7 -74.2    18    -5 A
 7  2013     1     1     6 EWR    FLL   N516JB  B6      Newa~  40.7 -74.2    18    -5 A
 8  2013     1     1     6 LGA    IAD   N829AS  EV      La G~  40.8 -73.9    22    -5 A
 9  2013     1     1     6 JFK    MCO   N593JB  B6      John~  40.6 -73.8    13    -5 A
10  2013     1     1     6 LGA    ORD   N3ALAA  AA      La G~  40.8 -73.9    22    -5 A
# ... with 336,766 more rows, and 1 more variable: tzone <chr>
```

Другие реализации

dplyr	base::merge
inner_join(x, y, by = "z")	merge(x, y)
left_join(x, y, by = "z")	merge(x, y, all.x = TRUE)
right_join(x, y, by = "z")	merge(x, y, all.y = TRUE)
full_join(x, y, by = "z")	merge(x, y, all.x = TRUE, all.y = TRUE)

dplyr	SQL
inner_join(x, y, by = "z")	SELECT * FROM x INNER JOIN y USING (z) SELECT * FROM x INNER JOIN y ON x.a = y.b.
left_join(x, y, by = "z")	SELECT * FROM x LEFT OUTER JOIN y USING (z) SELECT * FROM x LEFT OUTER JOIN y ON x.a = y.b.
right_join(x, y, by = "z")	SELECT * FROM x RIGHT OUTER JOIN y USING (z) SELECT * FROM x RIGHT OUTER JOIN y ON x.a = y.b.
full_join(x, y, by = "z")	SELECT * FROM x FULL OUTER JOIN y USING (z) SELECT * FROM x FULL OUTER JOIN y ON x.a = y.b.

Filtering Joins (фильтрующие соединения)

semi_join(x, y) сохраняет все значения в x, для которых есть совпадения в y.

• anti_join(x, y) опускает все значения в x, для которых есть совпадения в y

Filtering Joins

```
top_dest <- flights %>%
  count(dest, sort = TRUE) %>%
  head(10)
```

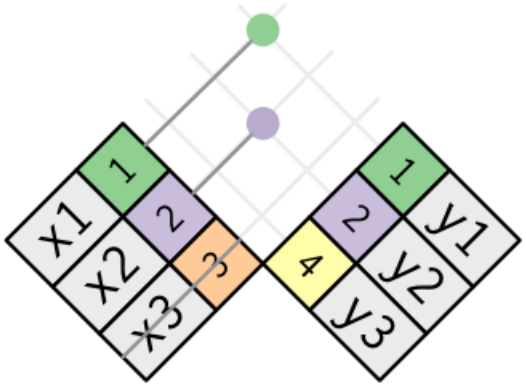
```
> top_dest
# A tibble: 10 x 2
   dest      n
   <chr> <int>
 1 ORD   17283
 2 ATL   17215
 3 LAX   16174
 4 BOS   15508
 5 MCO   14082
 6 CLT   14064
 7 SFO   13331
 8 FLL   12055
 9 MIA   11728
10 DCA    9705
```

```
flights %>%
  filter(dest %in% top_dest$dest)
```

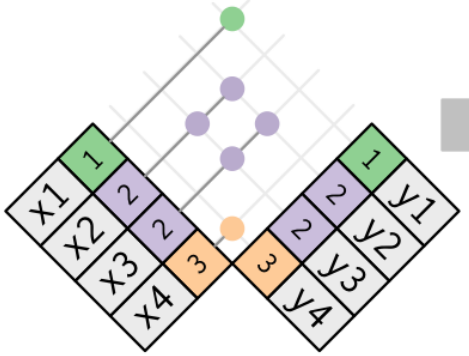
```
# A tibble: 141,145 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl>
 1  2013     1     1      542            540         2      923            850        33
 2  2013     1     1      554            600        -6      812            837       -25
 3  2013     1     1      554            558        -4      740            728        12
 4  2013     1     1      555            600        -5      913            854        19
 5  2013     1     1      557            600        -3      838            846        -8
 6  2013     1     1      558            600        -2      753            745         8
 7  2013     1     1      558            600        -2      924            917         7
 8  2013     1     1      558            600        -2      923            937       -14
 9  2013     1     1      559            559         0      702            706        -4
10  2013     1     1      600            600         0      851            858        -7
# ... with 141,135 more rows, and 10 more variables: carrier <chr>, flight <int>,
#   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
#   minute <dbl>, time_hour <dttm>
```

```
flights %>%
  semi_join(top_dest)
```

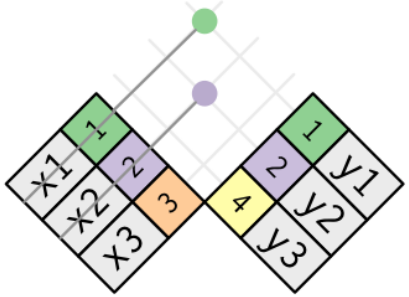
Filtering Joins



key	val_x
1	x1
2	x2



key	val_x
1	x1
2	x2
2	x3
3	x4



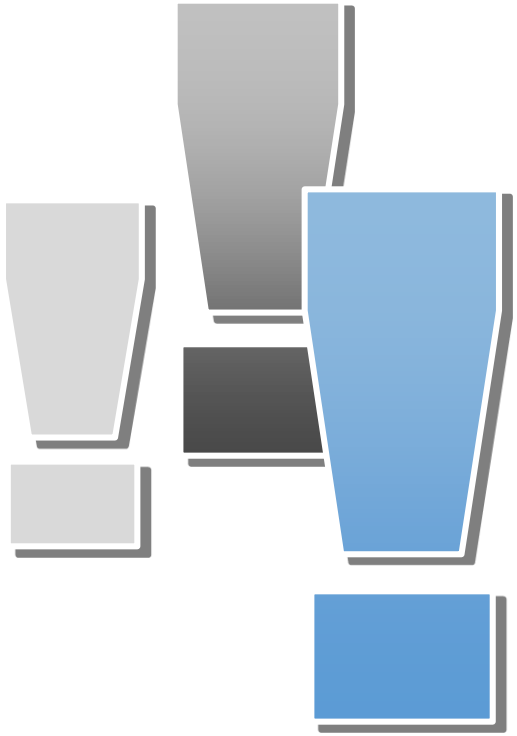
key	val_x
3	x3

Filtering Joins

```
flights %>%
  anti_join(planes, by = "tailnum") %>%
  count(tailnum, sort = TRUE)
```

```
# A tibble: 722 x 2
   tailnum     n
   <chr>   <int>
 1 NA       2512
 2 N725MQ    575
 3 N722MQ    513
 4 N723MQ    507
 5 N713MQ    483
 6 N735MQ    396
 7 N0EGMQ    371
 8 N534MQ    364
 9 N542MQ    363
10 N531MQ    349
# ... with 712 more rows
```

Спасибо за внимание!



Шевцов Василий Викторович

shevtsov_vv@rudn.university
+7(903)144-53-57