Data wrangling with dplyr

Анастасия Миллер, июль 2019

Хорошие данные

```
df <- read_csv('../data/fics_simple.csv')
```

Pipe operator %>%

Базовое использование:

%>% подставляет то,
что у него слева, первым аргументом в вызов справа:

- `f(x)` – то же, что и `x %>% f`
- `f(x, y)` – то же, что и `x %>% f(y)`
- `x %>% f %>% g %>% h` – то же, что и `h(g(f(x)))`

Pipe operator %>%

Базовое использование:

%>% подставляет то,
что у него **слева**, первым аргументом в вызов **справа**:

- `f(x)` – то же, что и `x %>% f`
- `f(x, y)` – то же, что и `x %>% f(y)`
- `x %>% f %>% g %>% h` – то же, что и `h(g(f(x)))`

Pipe operator %>%

Базовое использование:

%>% подставляет то,
что у него слева, первым аргументом в вызов справа:

- `f(x)` – то же, что и `x %>% f`
- `f(x, y)` – то же, что и `x %>% f(y)`
- `x %>% f %>% g %>% h` – то же, что и `h(g(f(x)))`

Pipe operator %>%

С указанием места:

- `x %>% f(y, .)` – то же, что и `f(y, x)`
- `x %>% f(y, z = .)` – то же, что и `f(y, z = x)`

Вопросы к данным

```
library(tidyr)
library(lubridate)
library(dplyr)
```

- Сколько всего фанфиков в нашем наборе?

Вопросы к данным

```
library(tidyr)
library(lubridate)
library(dplyr)
```

- Сколько всего фанфиков в нашем наборе?

```
df %>% nrow()
```

```
[1] 15824
```

Фильтрация: `filter` и `select`

- `filter` позволяет выбрать ряды, соответствующие некоторому условию (булевой маске):

```
df %>% filter(size_cat == 'small') %>% nrow()
```

```
## [1] 12413
```

- `select` позволяет выбрать колонки:

```
df %>% select(id, starts_with('size'))
```

```
       id size_cat size_kb
1  113637 medium       103
2  133114 small         12
3   82309 medium       157
…       … …              …
```

Вопросы к данным

- Сколько фанфиков имеют размер больше 100Кб?
- Когда опубликован самый первый фанфик? И что это был за фанфик?

Вопросы к данным

- Сколько фанфиков имеют размер больше 100Кб?

```
df %>% filter(size_kb > 100) %>% nrow()
```

```
[1] 2151
```

- Когда опубликован самый первый фанфик? И что это был за фанфик?

```
df %>% filter(published == min(published))
   %>% select(id, title, published)
```

```
# A tibble: 25 x 3
   id title                                                published
 1  6 Неприятности у Гарри                                2005-03-30
 2 16 Единство в смерти                                   2005-03-30
 3  2 Ни что не меняется так часто, как прошлое...        2005-03-30
 …  … …                                                   …
```

Разбор решений

Когда опубликован самый первый фанфик?
И что это был за фанфик?

- Через фильтр:
  ```
  df %>% filter(published == min(published))
  ```

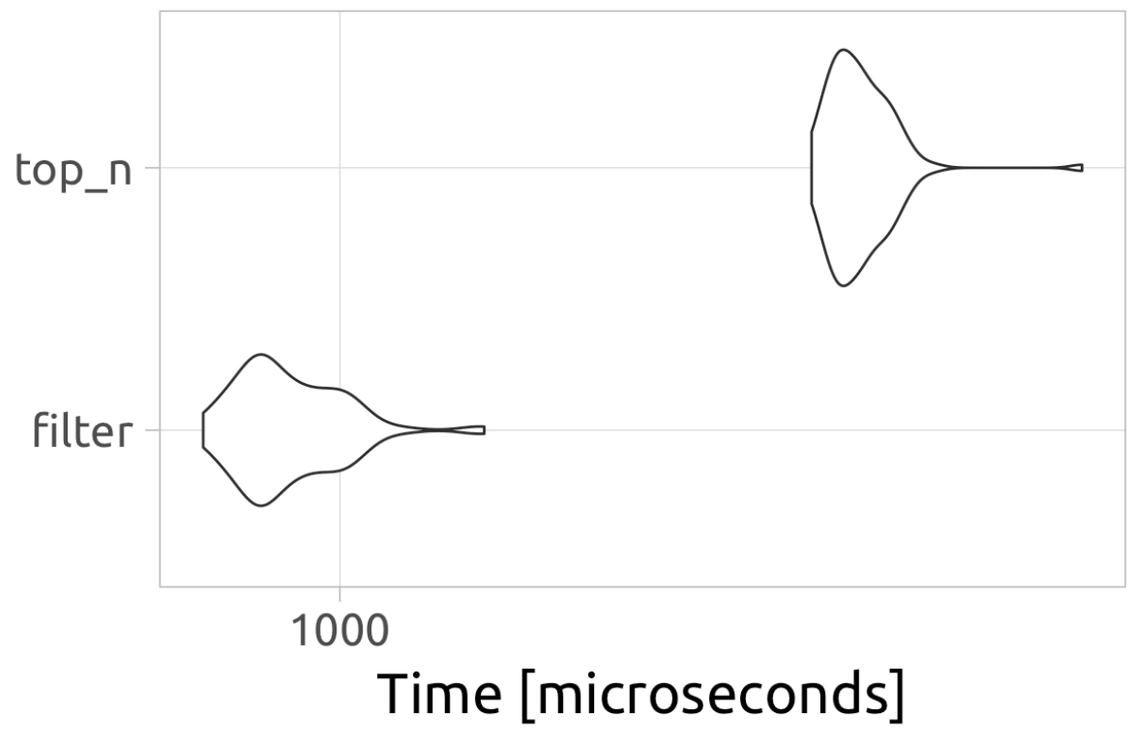
- Через `top_n`:
  ```
  df %>% top_n(1, published)
  ```

Какое из решений быстрее?

Разбор решений

Какое из решений быстрее?

```
microbenchmark(
  'filter' = {df %>% filter(published == max(published))},
  'top_n' = {df %>% top_n(1, published)}
) %>% autoplot()
```



Изменение данных: `mutate` и `arrange`

`mutate` добавляет новые переменные в набор, сохраняя старые:

```
df %>% filter(id %in% c(119387, 46832, 119034, 109013))
   %>% select(id, title, translated)
   %>% mutate(
    orig_lang = ifelse(translated, 'Unknown', 'Russian')
)
```

```
# A tibble: 4 x 4
     id title                    translated orig_lang
  <dbl> <chr>                    <lgl>      <chr>
1 109013 Мысли и возможности     FALSE      Russian
2 119034 Всегда                  TRUE       Unknown
3  46832 Томас                   FALSE      Russian
4 119387 Одно ветреное утро      TRUE       Unknown
```

Изменение данных: `mutate` и `arrange`

`transmute` добавляет новые переменные в набор, удаляя старые:

```
df %>% filter(id %in% c(119387, 46832, 119034, 109013))
    %>% select(id, title, translated)
    %>% transmute(
     orig_lang = ifelse(translated, 'Unknown', 'Russian')
)
```

```
# A tibble: 4 x 1
  orig_lang
  <chr>
1 Russian
2 Unknown
3 Russian
4 Unknown
```

Изменение данных: `mutate` и `arrange`

`arrange` сортирует по переменной:

```
df %>% select(title, size_kb, published)
   %>% arrange(size_kb, published)
```

```
# A tibble: 16,196 x 3
   title                size_kb published
   <chr>                  <int> <date>
 1 Все делают это.            0 2011-07-15
 2 Гусеница                   0 2014-05-02
 3 Без названия.              1 2005-03-30
```

Изменение данных: `mutate` и `arrange`

`desc` меняет направление сортировки:

```
df %>% select(title, size_kb, published)
    %>% arrange(published, desc(size_kb))
```

```
# A tibble: 16,196 x 3
   title                                        size_kb published
   <chr>                                          <int> <date>
 1 Гарри Поттер и Наследники Слизерина             1457 2005-03-30
 2 Гарри Поттер и Обряд Защиты Рода                 477 2005-03-30
 3 Гостья из Шармбатона                             247 2005-03-30
```

Вопросы к данным

- Какой фанфик писался дольше всего? Сколько времени заняло написание?

- Из тех фанфиков, которые писались больше года, какой был раньше всех начат и когда это случилось?

- Когда опубликовали первый переводной фанфик?

Полезно знать:

```
library(lubridate)
df$published %>% class()
[1] "character"
ymd(df$published) %>% class()
[1] "Date"
ymd('2019-02-15') - ymd('2019-01-23')
Time difference of 23 days
ymd('2019-02-15') - ymd('2019-01-23') > days(15)
[1] TRUE
```

- Сколько времени прошло от открытия сайта до завершения пятого опубликованного перевода?

Вопросы к данным

- Какой фанфик писался дольше всего? Сколько времени заняло написание?

```
df %>% mutate(
  published = ymd(published),
  last_update = ymd(last_update),
  time_to_complete = last_update - published
) %>% filter(
  time_to_complete == max(time_to_complete)
)%>% select(
  title, published, last_update, time_to_complete
)
```

```
     title   published last_update time_to_complete
1 Наследник  2005-04-06  2018-02-23        4706 days
```

Вопросы к данным

- Из тех фанфиков, которые писались больше года, какой был раньше всех начат и когда это случилось?

```
df %>% mutate(
  published = ymd(published),
  last_update = ymd(last_update),
  time_to_complete = last_update - published
) %>% filter(
  time_to_complete > years(1)
) %>% filter(
  published == min(published)
)
```

```
  id                 title  published last_update time_to_complete
1 679          Калейдоскоп  2005-04-04  2006-12-24          629 days
2  52 Превратности Судьбы  2005-04-04  2006-11-26          601 days
```

Вопросы к данным

- Когда опубликовали первый переводной фанфик?

```
df %>% filter(translated)
    %>% pull(published) %>% min()
```

```
[1] "2005-03-30"
```

- Сколько времени прошло от открытия сайта до завершения пятого опубликованного перевода?

```
opening_day <- ymd(min(df$published))
fifth_translation_complete <- df %>%
  filter(translated) %>% arrange(published) %>%
  slice(5) %>% pull(last_update) %>% ymd()
fifth_translation_complete - opening_day
```

```
Time difference of 240 days
```

Агрегация: `group_by` и `summarise`

`summarise` считает статистики:

```
df %>% summarise(
  mean_size = mean(size_kb),
  mean_ttc = mean(ymd(last_update) - ymd(published))
)
```

```
  mean_size       mean_ttc
1  71.03501 42.94161 days
```

Агрегация: `group_by` и `summarise`

`group_by` позволяет считать внутри группы, а не во всём наборе данных:

```
df %>% group_by(
  size_cat
) %>% summarise(
  mean_size = mean(size_kb),
  mean_ttc = mean(ymd(last_update) - ymd(published)),
  median_ttc = median(ymd(last_update) - ymd(published))
)
```

```
  size_cat mean_size         mean_ttc median_ttc
1    large  588.1886 347.928778 days   113 days
2   medium  106.2264  79.485418 days     4 days
3    small   15.3304   7.735434 days     0 days
```

Вопросы к данным

- Здесь есть есть авторские произведения и переводы. Сколько авторских, сколько переводов?

- Как часто появляются новые фанфики? Посчитайте, сколько в среднем фанфиков появляется в месяц.

- Отличается ли среднемесячное количество новых фанфиков, рассчитанное за всё время существования сайта, от рассчитанного за последние пять лет?

- Отличается ли среднемесячное количество публикаций для фанфиков разных рейтингов?

Разбор решений

Здесь есть есть авторские произведения и переводы. Сколько авторских, сколько переводов?

- Через `group_by`:
```
df %>% group_by(translated) %>% count()
```

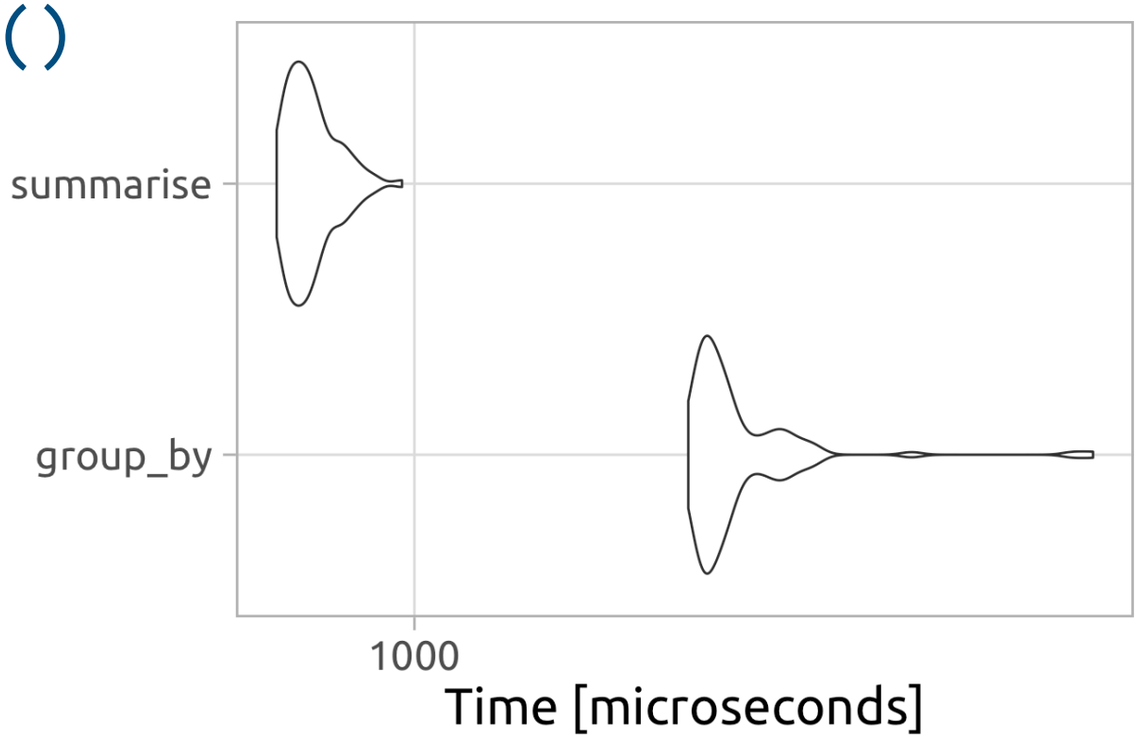
- Через `summarise`:
```
df %>% summarise(
  authored = sum(!translated),
  translated = sum(translated)
)
```

Какое из решений быстрее?

Разбор решений

Какое из решений быстрее?

```
microbenchmark(
   'group_by' = {df %>% group_by(translated) %>% count()},
   'summarise' = {df %>% summarise(
      authored = sum(!translated),
      translated = sum(translated))}
) %>% autoplot()
```



Вопросы к данным

- Как часто появляются новые фанфики? Посчитайте, сколько в среднем фанфиков появляется в месяц.

```
df %>% mutate(
    published_month = floor_date(published, "months")
  ) %>%
  group_by(published_month) %>%
    summarise(n = length(unique(id)))  %>%
  ungroup() %>%
  summarise(monthly_rate = mean(n))
```

```
  monthly_rate
1      94.75449
```

Вопросы к данным

- Отличается ли среднемесячное количество новых фанфиков, рассчитанное за всё время существования сайта, от рассчитанного за последние пять лет?

```
df %>% mutate(
    published_month = floor_date(published, "months")
  ) %>%
  group_by(published_month) %>%
    summarise(n = length(unique(id))) %>%
  ungroup() %>%
  summarise(
    monthly_rate_overall = mean(n),
    monthly_rate_recent = mean(
      n[published_month > today() - years(5)]
    )
  )
```

```
# A tibble: 1 x 2
  monthly_rate_overall monthly_rate_recent
                 <dbl>                <dbl>
1             94.75449             149.5167
```

Вопросы к данным

- Отличается ли среднемесячное количество публикаций для фанфиков разных рейтингов?

```
df %>% mutate(
  rating = ifelse(rating == "G", "General", rating),
  published_month = floor_date(published, "months")
  ) %>%
  group_by(published_month, rating) %>%
    summarise(n = length(unique(id))) %>%
  ungroup() %>%
  group_by(rating) %>%
    summarise(monthly_rate = mean(n))
```

```
# A tibble: 5 x 2
  rating    monthly_rate
  <chr>            <dbl>
1 General        28.17333
2 NC-17          15.96104
3 PG-13          33.95000
4 R              20.00000
5 не указан      14.66667
```

Бонусное задание

- Отличается ли частота публикаций для фанфиков разных рейтингов?

- Постройте доверительные интервалы для частоты в разных рейтингах в том же пайплайне

- Является ли различие статистически значимым?

Плохие данные

```
Sys.setlocale('LC_CTYPE', 'UTF-8')
library(jsonlite)
df <- stream_in(
    file('../data/fics.jsl'),
    simplifyMatrix=FALSE
)
```

Работа со списками: unnest

- `unnest` разворачивает список, создавая для каждого элемента новый ряд в данных:

```
df %>% slice(2, 5, 8)
   %>% select(id, title, genre)
```

```
     id                    title         genre
1 133114                  Умираем  Ангст, Драма
2 132668 Ни больше, ни меньше             Ангст
3 101366                Triangles  Драма, Ангст
```

Работа со списками: unnest

- `unnest` разворачивает список, создавая для каждого элемента новый ряд в данных:

```
df %>% slice(2, 5, 8)
    %>% select(id, title, genre) %>% unnest(genre)
```

```
      id                    title genre
1 133114                  Умираем  Ангст
2 133114                  Умираем  Драма
3 132668 Ни больше, ни меньше  Ангст
4 101366                Triangles  Драма
5 101366                Triangles  Ангст
```

Вопросы к данным

- Каков самый популярный жанр?

- Каков был самый популярный жанр в каждом из последних пяти лет?
  Жанр фанфика учитывается в том году, в котором фанфик был опубликован

- С какими жанрами чаще всего сочетается AU (alternative universe)? сложно

Вопросы к данным

- Каков самый популярный жанр?

```
df %>% unnest(genre)
   %>% group_by(genre)
   %>% summarise(n = length(unique(id)))
   %>% arrange(desc(n)) %>% slice(1:5)
```

```
# A tibble: 5 x 2
  genre          n
  <chr>      <int>
1 Романтика   5801
2 Драма       4575
3 Юмор        3364
4 Ангст       2674
5 Общий       1940
```

Вопросы к данным

- Каков был самый популярный жанр в каждом из последних пяти лет?

```
df %>% filter(year(published) > year(today()) - 5)
    %>% unnest(genre)
    %>% group_by(year(published), genre)
    %>% summarise(n = length(unique(id)))
    %>% top_n(1, wt = n)
```

```
# A tibble: 5 x 3
# Groups:   year(published) [5]
  `year(published)` genre          n
              <dbl> <chr>      <int>
1              2015 Романтика    589
2              2016 Романтика    557
3              2017 Драма        579
4              2018 Романтика    493
5              2019 Романтика    335
```

Вопросы к данным

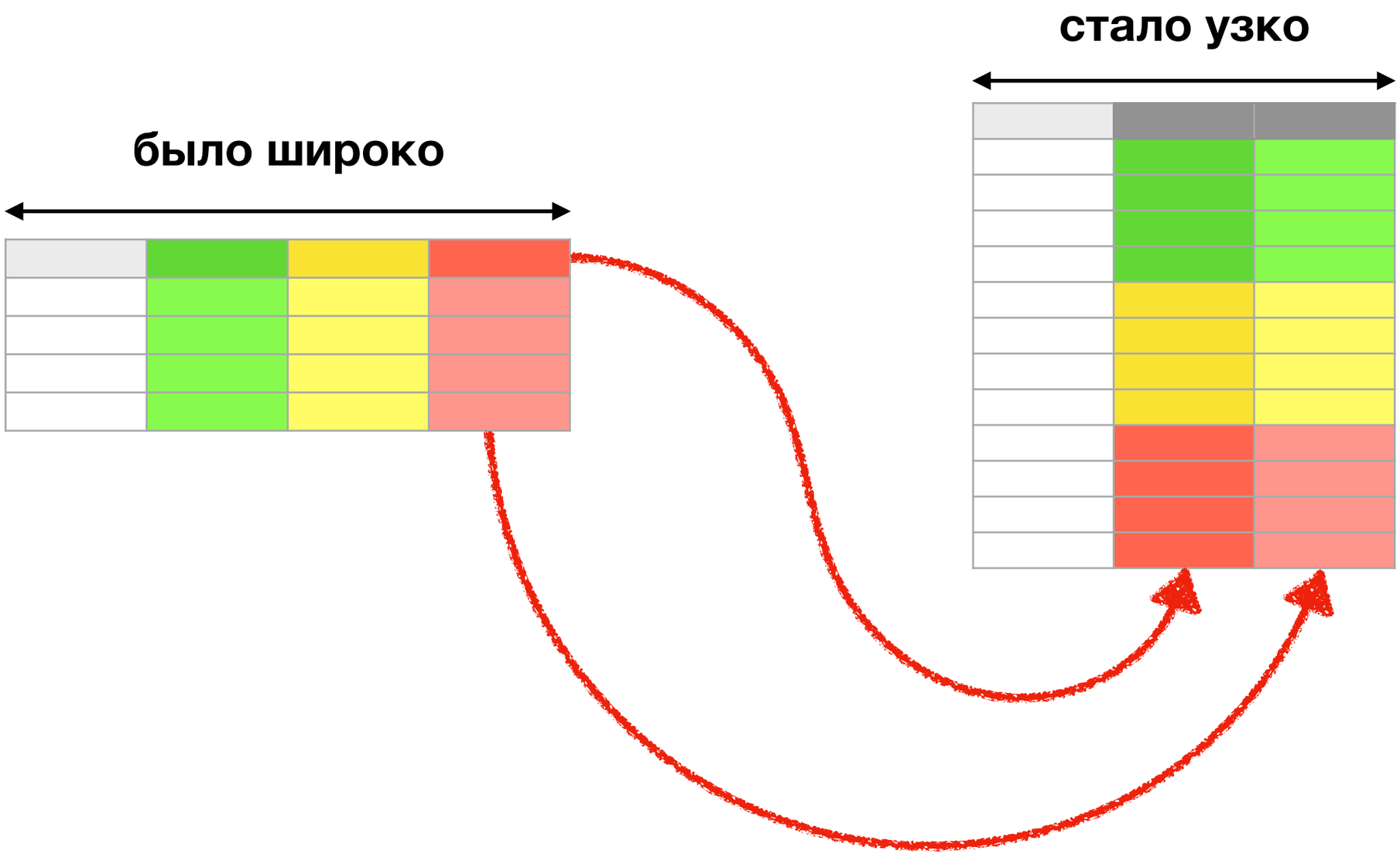
- С какими жанрами чаще всего сочетается PWP (porn without plot)?

```
df %>% filter(
      genre %>% sapply(function(x) "PWP" %in% x)
 ) %>% unnest(genre)
    %>% filter(genre != "PWP")
    %>% group_by(genre)
    %>% summarise(n = length(unique(id)))
    %>% top_n(5) %>% arrange(desc(n))
```

```
# A tibble: 5 x 2
  genre          n
  <chr>      <int>
1 Романтика    292
2 Юмор         186
3 Драма        100
4 Драббл        63
5 Флафф         60
```

Работа с untidy data: gather и spread

- `gather` собирает данные:



Работа с untidy data: gather и spread

- gather собирает данные:

```
df %>% select(id, published, last_update)
    %>% slice(1:3)
```

```
     id  published last_update
1 133114 2019-07-27  2019-07-27
2 132668 2019-07-14  2019-07-14
3 133246 2019-07-29  2019-07-29
```

Работа с untidy data: gather и spread

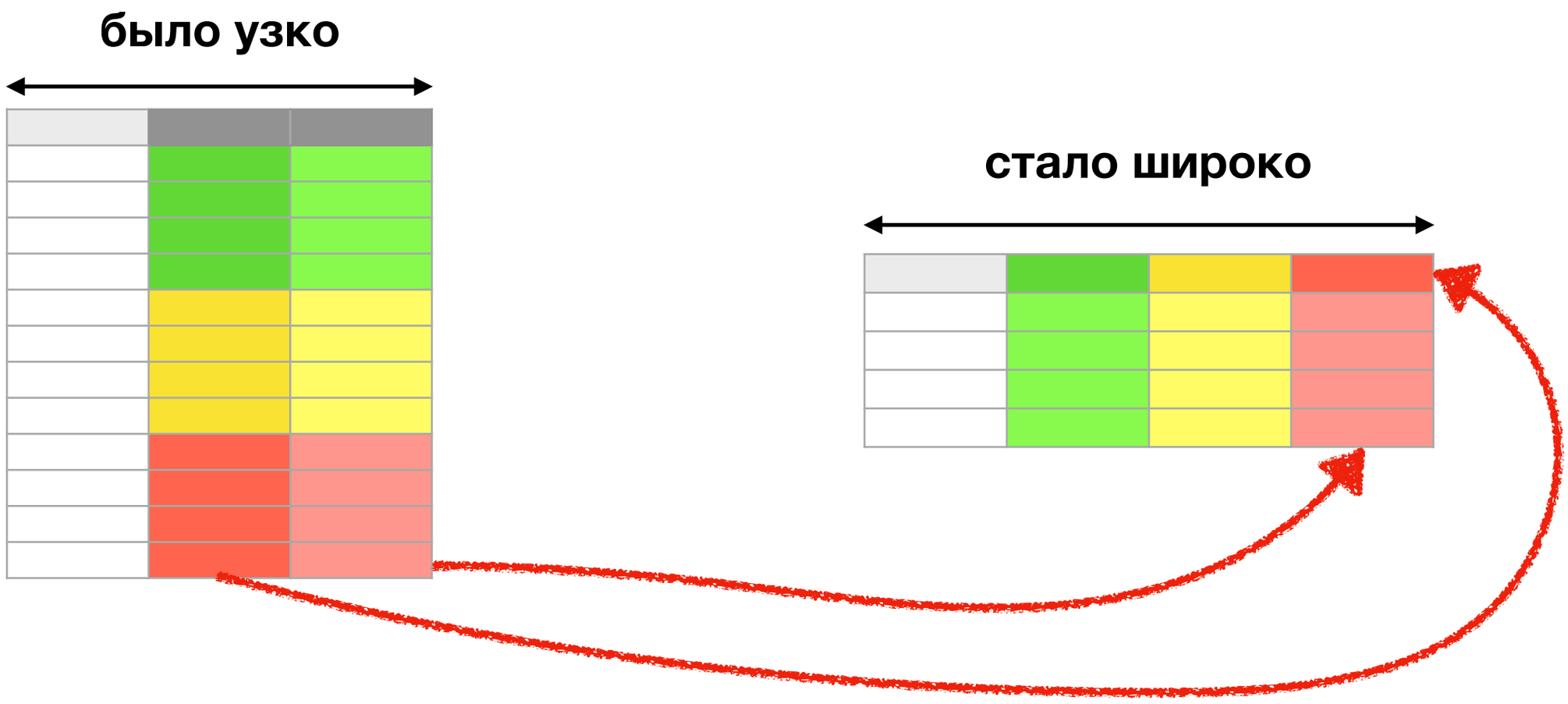
- gather собирает данные

```
df %>% select(id, published, last_update)
    %>% slice(1:3)
    %>% gather("action", "date", published, last_update)
```

```
     id       action       date
1 133114    published 2019-07-27
2 132668    published 2019-07-14
3 133246    published 2019-07-29
4 133114 last_update 2019-07-27
5 132668 last_update 2019-07-14
6 133246 last_update 2019-07-29
```

Работа с untidy data: gather и spread

- spread распределяет данные в широкую таблицу:



Работа с untidy data: gather и spread

- spread распределяет данные в широкую таблицу:

```
df %>% filter(year(ymd(published)) > 2016)
   %>% count(rating, year = year(published))
```

Работа с untidy data: gather и spread

- spread распределяет данные в широкую таблицу:

```
df %>% filter(year(ymd(published)) > 2016)
   %>% count(rating, year = year(published))
   %>% spread(year, n)
```

Вопросы к данным

- Соберите отчёт по количеству публикаций в пяти наиболее популярных жанрах за последние 10 лет. Каждый год — отдельная колонка.

- Посчитайте среднемесячное количество публикаций различного рейтинга. Учтите, что в некоторых месяцах могло не быть фанфиков какого-то рейтинга.

Вопросы к данным

- Соберите отчёт по количеству публикаций в пяти наиболее популярных жанрах за последние 7 лет. Каждый год — отдельная колонка.

```
top_5_genres <- df %>%
  filter(year(published) >= year(today()) - 7) %>%
  unnest(genre) %>% count(genre) %>%
  top_n(5, n) %>% pull(genre)

df %>% unnest(genre)
   %>% filter(
      (genre %in% top_5_genres) &
      (year(published) > year(today()) - 7))
   %>% count(genre, pub_year = year(published))
   %>% spread(pub_year, n)
```

Вопросы к данным

- Посчитайте среднемесячное количество публикаций различного рейтинга. Учтите, что в некоторых месяцах могло не быть фанфиков какого-то рейтинга.

```
df %>% select(id, published, rating)
%>% count(pub_month = ymd(published)
%>% floor_date("months"), rating)
%>% spread(pub_month, n, fill = 0)
%>% gather("pub_month", "n", 2:ncol(.))
%>% group_by(rating)
%>% summarise(mean(n))
```

Вопросы не к данным