RUDN university

5100

Программирование в среде R

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

dplyr()

Данные

https://www.transtats.bts.gov/DatabaseInfo.asp?DB_ID=120&Link=0

Эта база данных содержит информацию о запланированном и фактическом времени вылета и прилета сертифицированных американских авиаперевозчиков, на долю которых приходится не менее одного процента внутренних регулярных доходов пассажиров. Эти данные собираются управлением информации авиакомпаний, Бюро транспортной статистики (BTS).

```
install.packages("nycflights13")
library(nycflights13)
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 dest  air_time distance  hour minute
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>    <int> <chr>   <chr>  <chr>    <dbl>    <dbl> <dbl>  <dbl>
 1  2013     1     1      517            515         2      830            819        11 UA        1545 N14228  EWR    IAH        227     1400     5     15
 2  2013     1     1      533            529         4      850            830        20 UA        1714 N24211  LGA    IAH        227     1416     5     29
 3  2013     1     1      542            540         2      923            850        33 AA        1141 N619AA  JFK    MIA        160     1089     5     40
 4  2013     1     1      544            545        -1     1004           1022       -18 B6         725 N804JB  JFK    BQN        183     1576     5     45
 5  2013     1     1      554            600        -6      812            837       -25 DL         461 N668DN  LGA    ATL        116      762     6      0
 6  2013     1     1      554            558        -4      740            728        12 UA        1696 N39463  EWR    ORD        150      719     5     58
 7  2013     1     1      555            600        -5      913            854        19 B6         507 N516JB  EWR    FLL        158     1065     6      0
 8  2013     1     1      557            600        -3      709            723       -14 EV        5708 N829AS  LGA    IAD         53      229     6      0
 9  2013     1     1      557            600        -3      838            846        -8 B6          79 N593JB  JFK    MCO        140      944     6      0
10  2013     1     1      558            600        -2      753            745         8 AA         301 N3ALAA  LGA    ORD        138      733     6      0
# ... with 336,766 more rows, and 1 more variable: time_hour <dttm>
```

RUDN university

5100

Манипулирование с данными

- filter() – выбор наблюдений по их значениям
- arrange() – перестановка строк
- select() – выбор переменных по их именам
- mutate() – создание новых переменных с использованием функций существующих переменных
- summarize() – сведение нескольких значений в одно итоговое
- group_by() – группировка данных

Синтаксис:

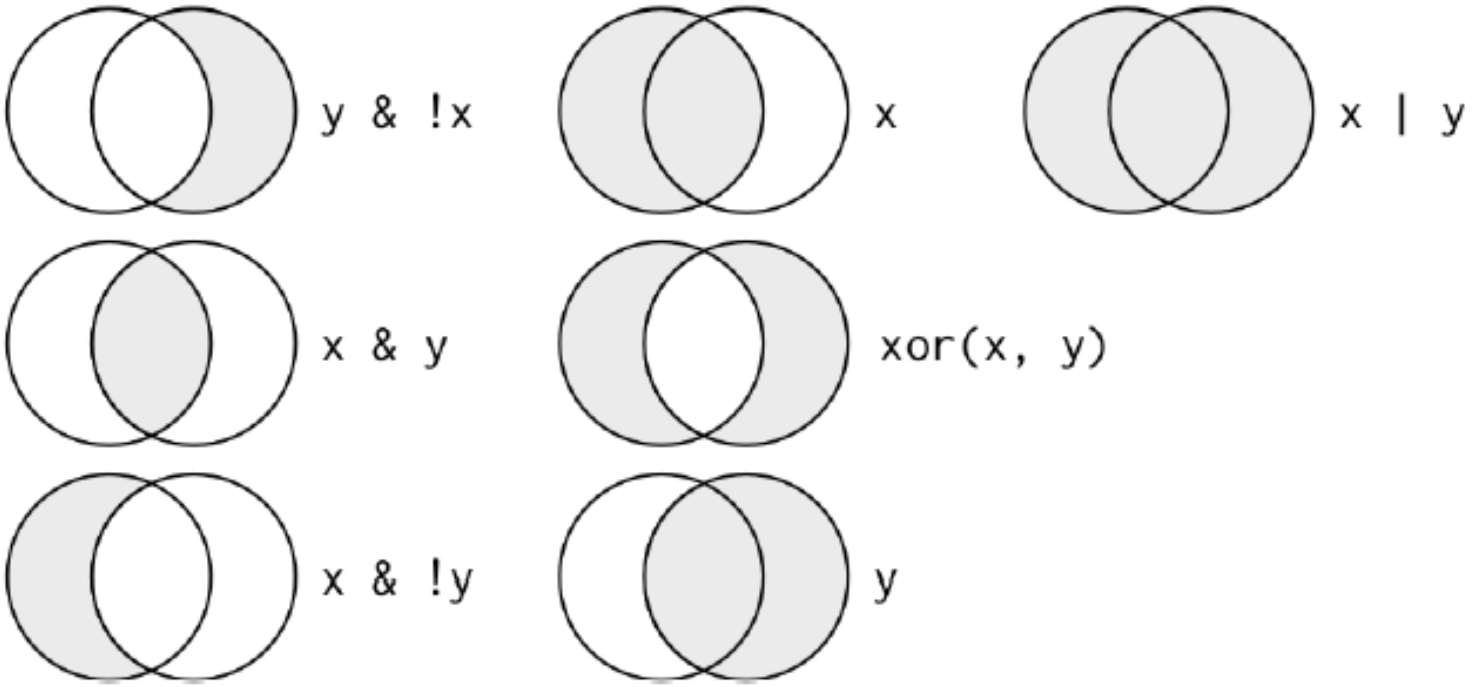
- Первый аргумент – фрейм данных
- Последующие аргументы – имена переменных (без кавычек), описывают действия, которые должны быть выполнены по отношению к фрейму данных
- Результатом является новый фрейм данных

filter()

```
filter(flights, month == 1, day == 1)
#> # A tibble: 842 × 19
#>    year month   day dep_time sched_dep_time dep_delay
#>   <int> <int> <int>    <int>          <int>     <dbl>
#> 1  2013     1     1      517            515         2
#> 2  2013     1     1      533            529         4
#> 3  2013     1     1      542            540         2
#> 4  2013     1     1      544            545        -1
#> 5  2013     1     1      554            600        -6
#> 6  2013     1     1      554            558        -4
#> # ... with 836 more rows,
#> #   arr_time <int>, sched_
#> #   carrier <chr>, flight
#> #   dest <chr>, air_time <
#> #   minute <dbl>, time_hou
```

```
(dec25 <- filter(flights, month == 12, day == 25))
#> # A tibble: 719 × 19
#>    year month   day dep_time sched_dep_time dep_delay
#>   <int> <int> <int>    <int>          <int>     <dbl>
#> 1  2013    12    25      456            500        -4
#> 2  2013    12    25      524            515         9
#> 3  2013    12    25      542            540         2
#> 4  2013    12    25      546            550        -4
#> 5  2013    12    25      556            600        -4
#> 6  2013    12    25      557            600        -3
#> # ... with 713 more rows, and 13 more variables:
#> #   arr_time <int>, sched_arr_time <int>, arr_delay <dbl>,
#> #   carrier <chr>, flight <int>, tailnum <chr>,origin <chr>,
#> #   dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
#> #   minute <dbl>, time_hour <dttm>
```

filter() Логические операторы



```
> filter(flights, month %in% c(11, 12))
# A tibble: 55,403 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>
2013	11	1	5	2359	6	352	345	7	B6	745	N568JB	JFK
2013	11	1	35	2250	105	123	2356	87	B6	1816	N353JB	JFK
2013	11	1	455	500	-5	641	651	-10	US	1895	N192UW	EWR
2013	11	1	539	545	-6	856	827	29	UA	1714	N38727	LGA
2013	11	1	542	545	-3	831	855	-24	AA	2243	N5CLAA	JFK
2013	11	1	549	600	-11	912	923	-11	UA	303	N595UA	JFK
2013	11	1	550	600	-10	705	659	6	US	2167	N748UW	LGA
2013	11	1	554	600	-6	659	701	-2	US	2134	N742PS	LGA
2013	11	1	554	600	-6	826	827	-1	DL	563	N912DE	LGA
2013	11	1	554	600	-6	749	751	-2	DL	731	N315NB	LGA

```
# ... with 55,393 more rows, and 6 more variables: dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
#   minute <dbl>, time_hour <dttm>
```

```
> filter(flights, !(arr_delay > 120 | dep_delay > 120))
# A tibble: 316,050 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>
2013	1	1	517	515	2	830	819	11	UA	1545	N14228	EWR
2013	1	1	533	529	4	850	830	20	UA	1714	N24211	LGA
2013	1	1	542	540	2	923	850	33	AA	1141	N619AA	JFK
2013	1	1	544	545	-1	1004	1022	-18	B6	725	N804JB	JFK
2013	1	1	554	600	-6	812	837	-25	DL	461	N668DN	LGA
2013	1	1	554	558	-4	740	728	12	UA	1696	N39463	EWR
2013	1	1	555	600	-5	913	854	19	B6	507	N516JB	EWR
2013	1	1	557	600	-3	709	723	-14	EV	5708	N829AS	LGA
2013	1	1	557	600	-3	838	846	-8	B6	79	N593JB	JFK
2013	1	1	558	600	-2	753	745	8	AA	301	N3ALAA	LGA

```
> filter(flights, arr_delay <= 120, dep_delay <= 120)
# A tibble: 316,050 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>
2013	1	1	517	515	2	830	819	11	UA	1545	N14228	EWR
2013	1	1	533	529	4	850	830	20	UA	1714	N24211	LGA
2013	1	1	542	540	2	923	850	33	AA	1141	N619AA	JFK
2013	1	1	544	545	-1	1004	1022	-18	B6	725	N804JB	JFK
2013	1	1	554	600	-6	812	837	-25	DL	461	N668DN	LGA
2013	1	1	554	558	-4	740	728	12	UA	1696	N39463	EWR
2013	1	1	555	600	-5	913	854	19	B6	507	N516JB	EWR
2013	1	1	557	600	-3	709	723	-14	EV	5708	N829AS	LGA
2013	1	1	557	600	-3	838	846	-8	B6	79	N593JB	JFK
2013	1	1	558	600	-2	753	745	8	AA	301	N3ALAA	LGA

```
# ... with 316,040 more rows, and 6 more variables: dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
#   minute <dbl>, time_hour <dttm>
```

NA

```
> NA > 5
[1] NA
> 10 == NA
[1] NA
> NA + 10
[1] NA
> NA / 2
[1] NA
> NA == NA
[1] NA
```

```
> df <- tibble(x = c(1, NA, 3))
> filter(df, x > 1)
# A tibble: 1 x 1
      x
  <dbl>
1     3
> filter(df, is.na(x) | x > 1)
# A tibble: 2 x 1
      x
  <dbl>
1    NA
2     3
```

arrange()

```
> arrange(flights, year, month, day)
# 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>
```

```
> arrange(flights, desc(arr_delay))
# 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     9      641            900      1301     1242           1530      1272 HA         51 N384HA  JFK
 2  2013     6    15     1432           1935      1137     1607           2120      1127 MQ       3535 N504MQ  JFK
 3  2013     1    10     1121           1635      1126     1239           1810      1109 MQ       3695 N517MQ  EWR
 4  2013     9    20     1139           1845      1014     1457           2210      1007 AA        177 N338AA  JFK
 5  2013     7    22      845           1600      1005     1044           1815       989 MQ       3075 N665MQ  JFK
 6  2013     4    10     1100           1900       960     1342           2211       931 DL       2391 N959DL  JFK
 7  2013     3    17     2321            810       911      135           1020       915 DL       2119 N927DA  LGA
 8  2013     7    22     2257            759       898      121           1026       895 DL       2047 N6716C  LGA
 9  2013    12     5      756           1700       896     1058           2020       878 AA        172 N5DMAA  EWR
10  2013     5     3     1133           2055       878     1250           2215       875 MQ       3744 N523MQ  EWR
# ... with 336,766 more rows, and 6 more variables: dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
#   minute <dbl>, time_hour <dttm>
```

arrange()

```
> df <- tibble(x = c(5, 2, NA))
> arrange(df, x)
# A tibble: 3 x 1
      x
  <dbl>
1     2
2     5
3    NA
```

```
> arrange(df, desc(x))
# A tibble: 3 x 1
      x
  <dbl>
1     5
2     2
3    NA
```

select()

```
> # Select columns by name
> select(flights, year, month, day)
# A tibble: 336,776 x 3
    year month   day
   <int> <int> <int>
 1  2013     1     1
 2  2013     1     1
 3  2013     1     1
 4  2013     1     1
 5  2013     1     1
 6  2013     1     1
 7  2013     1     1
 8  2013     1     1
 9  2013     1     1
10  2013     1     1
# ... with 336,766 more rows
```

```
> # Select all columns between year and day (inclusive)
> select(flights, year:day)
# A tibble: 336,776 x 3
    year month   day
   <int> <int> <int>
 1  2013     1     1
 2  2013     1     1
 3  2013     1     1
 4  2013     1     1
 5  2013     1     1
 6  2013     1     1
 7  2013     1     1
 8  2013     1     1
 9  2013     1     1
10  2013     1     1
# ... with 336,766 more rows
```

```
> # Select all columns except those from year to day (inclusive)
> select(flights, -(year:day))
# A tibble: 336,776 x 16
   dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier flight tailnum origin dest  air_time
      <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>    <int> <chr>   <chr>  <chr>    <dbl>
 1      517            515         2      830            819        11 UA        1545 N14228  EWR    IAH        227
 2      533            529         4      850            830        20 UA        1714 N24211  LGA    IAH        227
 3      542            540         2      923            850        33 AA        1141 N619AA  JFK    MIA        160
 4      544            545        -1     1004           1022       -18 B6         725 N804JB  JFK    BQN        183
 5      554            600        -6      812            837       -25 DL         461 N668DN  LGA    ATL        116
 6      554            558        -4      740            728        12 UA        1696 N39463  EWR    ORD        150
 7      555            600        -5      913            854        19 B6         507 N516JB  EWR    FLL        158
 8      557            600        -3      709            723       -14 EV        5708 N829AS  LGA    IAD         53
 9      557            600        -3      838            846        -8 B6          79 N593JB  JFK    MCO        140
10      558            600        -2      753            745         8 AA         301 N3ALAA  LGA    ORD        138
# ... with 336,766 more rows, and 4 more variables: distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

select()

Дополнительные функции

- starts_with("abc") соответствует именам, начинающимся с “abc”
- ends_with("xyz") соответствует именам, заканчивающимся на “xyz”
- contains("ijk") соответствует именам, содержащим “ijk”
- matches("(.)\\1") выбирает переменные, соответствующие регулярному выражению
- num_range("x", 1:3) соответствует x1, x2, and x3

select()

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

```
> select(flights, time_hour, air_time, everything())
# A tibble: 336,776 x 19
   time_hour           air_time  year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay
   <dttm>                 <dbl> <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl>
 1 2013-01-01 05:00:00      227  2013     1     1      517            515         2      830            819        11
 2 2013-01-01 05:00:00      227  2013     1     1      533            529         4      850            830        20
 3 2013-01-01 05:00:00      160  2013     1     1      542            540         2      923            850        33
 4 2013-01-01 05:00:00      183  2013     1     1      544            545        -1     1004           1022       -18
 5 2013-01-01 06:00:00      116  2013     1     1      554            600        -6      812            837       -25
 6 2013-01-01 05:00:00      150  2013     1     1      554            558        -4      740            728        12
 7 2013-01-01 06:00:00      158  2013     1     1      555            600        -5      913            854        19
 8 2013-01-01 06:00:00       53  2013     1     1      557            600        -3      709            723       -14
 9 2013-01-01 06:00:00      140  2013     1     1      557            600        -3      838            846        -8
10 2013-01-01 06:00:00      138  2013     1     1      558            600        -2      753            745         8
# ... with 336,766 more rows, and 8 more variables: carrier <chr>, flight <int>, tailnum <chr>, origin <chr>,
#   dest <chr>, distance <dbl>, hour <dbl>, minute <dbl>
```

mutate()

```
flights_sml <- select(flights,
                      year:day,
                      ends_with("delay"),
                      distance,
                      air_time
)
mutate(flights_sml,
       gain = arr_delay - dep_delay,
       speed = distance / air_time * 60
)
```

```
# A tibble: 336,776 x 9
    year month   day dep_delay arr_delay distance air_time  gain speed
   <int> <int> <int>     <dbl>     <dbl>    <dbl>    <dbl> <dbl> <dbl>
 1  2013     1     1         2        11     1400      227     9  370.
 2  2013     1     1         4        20     1416      227    16  374.
 3  2013     1     1         2        33     1089      160    31  408.
 4  2013     1     1        -1       -18     1576      183   -17  517.
 5  2013     1     1        -6       -25      762      116   -19  394.
 6  2013     1     1        -4        12      719      150    16  288.
 7  2013     1     1        -5        19     1065      158    24  404.
 8  2013     1     1        -3       -14      229       53   -11  259.
 9  2013     1     1        -3        -8      944      140    -5  405.
10  2013     1     1        -2         8      733      138    10  319.
# ... with 336,766 more rows
```

mutate()

```
mutate(flights_sml,
       gain = arr_delay - dep_delay,
       hours = air_time / 60,
       gain_per_hour = gain / hours
)
```

```
# A tibble: 336,776 x 10
    year month   day dep_delay arr_delay distance air_time  gain hours gain_per_hour
   <int> <int> <int>     <dbl>     <dbl>    <dbl>    <dbl> <dbl> <dbl>         <dbl>
 1  2013     1     1         2        11     1400      227     9 3.78           2.38
 2  2013     1     1         4        20     1416      227    16 3.78           4.23
 3  2013     1     1         2        33     1089      160    31 2.67          11.6
 4  2013     1     1        -1       -18     1576      183   -17 3.05          -5.57
 5  2013     1     1        -6       -25      762      116   -19 1.93          -9.83
 6  2013     1     1        -4        12      719      150    16 2.5            6.4
 7  2013     1     1        -5        19     1065      158    24 2.63           9.11
 8  2013     1     1        -3       -14      229       53   -11 0.883        -12.5
 9  2013     1     1        -3        -8      944      140    -5 2.33          -2.14
10  2013     1     1        -2         8      733      138    10 2.3            4.35
# ... with 336,766 more rows
```

mutate()

```
transmute(flights,
          gain = arr_delay - dep_delay,
          hours = air_time / 60,
          gain_per_hour = gain / hours
)
```

```
# A tibble: 336,776 x 3
    gain hours gain_per_hour
   <dbl> <dbl>         <dbl>
 1     9 3.78           2.38
 2    16 3.78           4.23
 3    31 2.67          11.6
 4   -17 3.05          -5.57
 5   -19 1.93          -9.83
 6    16 2.5            6.4
 7    24 2.63           9.11
 8   -11 0.883        -12.5
 9    -5 2.33          -2.14
10    10 2.3            4.35
# ... with 336,766 more rows
```

функции mutate()

- Арифметические операторы +, -, *, /
- Модулярная арифметика %/% (целочисленное деление), %% (остаток)
- Смещения. Функции lead() lag() позволяют ссылаться на значения, отстоящие от заданного на указанное число позиций

```
> (x <- 1:10)
 [1]  1  2  3  4  5  6  7  8  9 10
> lag(x)
 [1] NA  1  2  3  4  5  6  7  8  9
> lead(x)
 [1]  2  3  4  5  6  7  8  9 10 NA
```

- Скользящие cumsum(), cumprod(), cummin(), cummax(), dplyr::cummean()

```
> x
 [1]  1  2  3  4  5  6  7  8  9 10
> cumsum(x)
 [1]  1  3  6 10 15 21 28 36 45 55
> cummean(x)
 [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5
```

- Логические операторы <, <=, >, >=, !=
- Ранжирование

```
> y <- c(1, 2, 2, NA, 3, 4)
> min_rank(y)
[1]  1  2  2 NA  4  5
> min_rank(desc(y))
[1]  5  3  3 NA  2  1
```

row_number(), dense_rank(), percent_rank(), cume_dist(), ntile().

```
> row_number(y)
[1]  1  2  3 NA  4  5
> dense_rank(y)
[1]  1  2  2 NA  3  4
> percent_rank(y)
[1] 0.00 0.25 0.25   NA 0.75 1.00
> cume_dist(y)
[1] 0.2 0.6 0.6  NA 0.8 1.0
```

summarize()

```
> summarize(flights, delay = mean(dep_delay, na.rm = TRUE))
# A tibble: 1 x 1
  delay
  <dbl>
1  12.6
```

```
> by_day <- group_by(flights, year, month, day)
> summarize(by_day, delay = mean(dep_delay, na.rm = TRUE))
# A tibble: 365 x 4
# Groups:   year, month [12]
    year month   day delay
   <int> <int> <int> <dbl>
 1  2013     1     1 11.5
 2  2013     1     2 13.9
 3  2013     1     3 11.0
 4  2013     1     4  8.95
 5  2013     1     5  5.73
 6  2013     1     6  7.15
 7  2013     1     7  5.42
 8  2013     1     8  2.55
 9  2013     1     9  2.28
10  2013     1    10  2.84
# ... with 355 more rows
```

конвейеры

```
delays <- flights %>%
  group_by(dest) %>%
  summarize(
    count = n(),
    dist = mean(distance, na.rm = TRUE),
    delay = mean(arr_delay, na.rm = TRUE)
  ) %>%
  filter(count > 20, dest != "HNL")
```

```
> delays
# A tibble: 96 x 4
   dest  count  dist delay
   <chr> <int> <dbl> <dbl>
 1 ABQ     254 1826   4.38
 2 ACK     265  199   4.85
 3 ALB     439  143  14.4
 4 ATL   17215  757. 11.3
 5 AUS    2439 1514.  6.02
 6 AVL     275  584.  8.00
 7 BDL     443  116   7.05
 8 BGR     375  378   8.03
 9 BHM     297  866. 16.9
10 BNA    6333  758. 11.8
# ... with 86 more rows
```

счетчик n()

```
not_cancelled <- flights %>%
                 filter(!is.na(dep_delay),
                 !is.na(arr_delay)
)
```

```
> not_cancelled
# A tibble: 327,346 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 327,336 more rows, and 6 more variables: dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
#   minute <dbl>, time_hour <dttm>
```

```
delays <- not_cancelled %>%
  group_by(tailnum) %>%
  summarize(
    delay = mean(arr_delay, na.rm = TRUE),
    n = n()
  )
```

```
> delays
# A tibble: 4,037 x 3
   tailnum  delay     n
   <chr>    <dbl> <int>
 1 D942DN   31.5      4
 2 N0EGMQ    9.98   352
 3 N10156   12.7    145
 4 N102UW    2.94    48
 5 N103US   -6.93    46
 6 N104UW    1.80    46
 7 N10575   20.7    269
 8 N105UW   -0.267   45
 9 N107US   -5.73    41
10 N108UW   -1.25    60
# ... with 4,027 more rows
```

Прочие функции

- средние mean(x), median(x)
- разброс sd(x), IQR(x), mad(x)
- ранжирование min(x), quantile(x, 0.25), max(x)
- порядковые min(x), quantile(x, 0.25), max(x)
- счетчики n(), sum(!is.na(x)), n_distinct(x)
- количество и доли логических значений sum(x > 10), mean(y == 0)

Примеры

Сколько авиарейсов отправлено до 5 часов утра?

```
not_cancelled %>%
group_by(year, month, day) %>%
summarize(n_early = sum(dep_time < 500))
```

```
# A tibble: 365 x 4
# Groups:   year, month [12]
    year month   day n_early
   <int> <int> <int>   <int>
 1  2013     1     1       0
 2  2013     1     2       3
 3  2013     1     3       4
 4  2013     1     4       3
 5  2013     1     5       3
 6  2013     1     6       2
 7  2013     1     7       2
 8  2013     1     8       1
 9  2013     1     9       3
10  2013     1    10       3
# ... with 355 more rows
```

Какова доля авиарейсов, задержанных более чем на один час?

```
not_cancelled %>%
group_by(year, month, day) %>%
summarize(hour_perc = mean(arr_delay > 60))
```

```
# A tibble: 365 x 4
# Groups:   year, month [12]
    year month   day hour_perc
   <int> <int> <int>     <dbl>
 1  2013     1     1    0.0722
 2  2013     1     2    0.0851
 3  2013     1     3    0.0567
 4  2013     1     4    0.0396
 5  2013     1     5    0.0349
 6  2013     1     6    0.0470
 7  2013     1     7    0.0333
 8  2013     1     8    0.0213
 9  2013     1     9    0.0202
10  2013     1    10    0.0183
# ... with 355 more rows
```

Регулярные выражения

```
x <- c("apple", "banana", "pear")
str_view(x, "an")
```

```
apple
banana
pear
```

```
str_view(x, ".a.")
```

```
apple
banana
pear
```

```
str_view(c("abc", "a.c", "bef"), "a\\.c")
```

```
abc
a.c
bef
```

```
str_view(x, "\\\\")
```

```
a\b
```

Регулярные выражения

```
x <- c("apple", "banana", "pear")
str_view(x, "^a")
```

```
apple
banana
pear
```

```
str_view(x, "a$")
```

```
apple
banana
pear
```

```
x <- c("apple pie", "apple", "apple cake")
str_view(x, "apple")
```

```
apple pie
apple
apple cake
```

```
str_view(x, "^apple$")
```

```
apple pie
apple
apple cake
```

Регулярные выражения

- \d совпадает с любой цифрой
- \s совпадает с любым пробельным символом
- [abc] совпадает с a, b, c.
- [^abc] совпадает с любым символом, кроме a, b, or c.

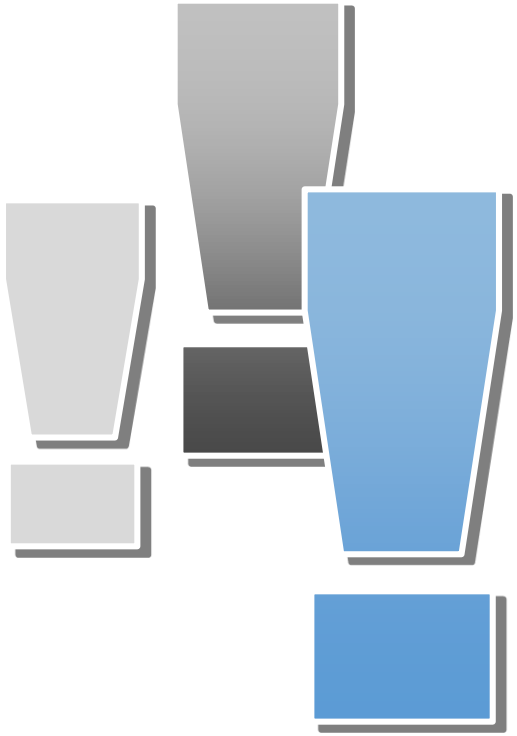
```
str_view(c("grey", "gray"), "gr(e|a)y")
```

Регулярные выражения

повторения
? – 0 или 1 раз
+ – 1 или более раз
* – 0 или более раз

количество повторений
{n} – ровно n раз
{n,} – n или более раз
{,m} – не более m раз
{n,m} – от n до m раз

Спасибо за внимание!



Шевцов Василий Викторович

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