Đáp án Bài tập thực hành 1
-Khám phá phân tích dữ liệu với R-

TS.Tô Đức Khánh

05/03/2024

```
library(nycflights13)
data(flights)
glimpse(flights)
```

```
## Rows: 336,776
## Columns: 19
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, ~
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558, 558, ~
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 600, 600, ~
## $ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -1, 0, ~
## $ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 849, 853,~
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 851, 856,~
## $ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7, -14, 31~
## $ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6", "AA", ~
## $ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301, 49, 71~
## $ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N39463", ~
## $ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA", "JFK~
## $ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IAD", "MCO~
## $ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149, 158, 3~
## $ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 733, 1028,~
## $ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6, 6, 6, ~
## $ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 0, 59, 0, 0, ~
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 05:00:~
```

1 Bài tập

Bài tập 1: Trong một quy trình duy nhất cho từng điều kiện, hãy tìm tất cả các chuyến bay đáp ứng điều kiện:

Đến nơi trễ từ hai giờ trở lên

```
flights |>
  filter(arr_delay >= 120) |>
  arrange(desc(arr_delay))
```

```
## # A tibble: 10,200 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     9      641            900      1301     1242           1530
```

```
## 2  2013     6    15     1432           1935      1137     1607           2120
## 3  2013     1    10     1121           1635      1126     1239           1810
## 4  2013     9    20     1139           1845      1014     1457           2210
## 5  2013     7    22      845           1600      1005     1044           1815
## # i 10,195 more rows
## # i 11 more variables: 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>
```

Bay tới Houston (IAH hoặc HOU)

```
flights |>
  filter(dest %in% c("IAH", "HOU"))
```

```
## # A tibble: 9,313 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     1      517            515         2      830            819
## 2  2013     1     1      533            529         4      850            830
## 3  2013     1     1      623            627        -4      933            932
## 4  2013     1     1      728            732        -4     1041           1038
## 5  2013     1     1      739            739         0     1104           1038
## # i 9,308 more rows
## # i 11 more variables: 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>
```

Được điều hành bởi United, American hoặc Delta

```
flights |>
  filter(carrier %in% c("UA", "DL", "AA"))
```

```
## # A tibble: 139,504 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     1      517            515         2      830            819
## 2  2013     1     1      533            529         4      850            830
## 3  2013     1     1      542            540         2      923            850
## 4  2013     1     1      554            600        -6      812            837
## 5  2013     1     1      554            558        -4      740            728
## # i 139,499 more rows
## # i 11 more variables: 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>
```

Khởi hành vào mùa hè (tháng 7, tháng 8, tháng 9)

```
flights |>
  filter(month %in% c(7, 8, 9))
```

```
## # A tibble: 86,326 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     7     1        1           2029       212      236           2359
## 2  2013     7     1        2           2359         3      344            344
## 3  2013     7     1       29           2245       104      151              1
## 4  2013     7     1       43           2130       193      322             14
## 5  2013     7     1       44           2150       174      300            100
```

```
## # i 86,321 more rows
## # i 11 more variables: 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>
```

Đến muộn hơn hai tiếng nhưng không cất cánh muộn

```
flights |>
  filter(arr_delay >= 120 & dep_delay <= 0)
```

```
## # A tibble: 29 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1    27     1419           1420        -1     1754           1550
## 2  2013    10     7     1350           1350         0     1736           1526
## 3  2013    10     7     1357           1359        -2     1858           1654
## 4  2013    10    16      657            700        -3     1258           1056
## 5  2013    11     1      658            700        -2     1329           1015
## # i 24 more rows
## # i 11 more variables: 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>
```

Bị trễ ít nhất một giờ, nhưng đã bù được hơn 30 phút trong thời gian chuyến bay

Đối với câu hỏi này, lưu ý rằng, về mặt lý thuyết, số phút cất cánh trễ của chuyến bay, `dep_delay` sẽ chính là số phút hạ cánh trễ của chuyến bay `arr_delay`. Tuy nhiên, trong thực tế, `arr_delay` có thể ngắn hơn hoặc dài hơn `dep_delay` do máy bay di chuyển nhanh hơn hoặc chậm hơn. Di chuyển nhanh hơn thì tới sớm hơn với lịch trình bị trễ. Do đó, "tới sớm hơn 30 phút so với lịch trình bị trễ" có nghĩa là `arr_delay` ngắn hơn 30 phút so với `dep_delay`.

```
flights |>
  filter(dep_delay >= 60 & dep_delay - arr_delay > 30)
```

```
## # A tibble: 1,844 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     1     2205           1720       285       46           2040
## 2  2013     1     1     2326           2130       116      131             18
## 3  2013     1     3     1503           1221       162     1803           1555
## 4  2013     1     3     1839           1700        99     2056           1950
## 5  2013     1     3     1850           1745        65     2148           2120
## # i 1,839 more rows
## # i 11 more variables: 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>
```

Bài tập 2: Sắp xếp `flights` để tìm chuyến bay có thời gian khởi hành trễ nhất. Tìm các chuyến bay khởi hành sớm nhất vào buổi sáng.

```
flights |>
  arrange(desc(dep_delay)) |>
  arrange(sched_dep_time) |>
  relocate(dep_delay, sched_dep_time)
```

```
## # A tibble: 336,776 x 19
##   dep_delay sched_dep_time  year month   day dep_time arr_time sched_arr_time
##       <dbl>          <int> <int> <int> <int>    <int>    <int>          <int>
```

```
## 1        NA            106  2013     7    27       NA       NA            245
## 2       188            500  2013     4    24      808     1008            640
## 3        61            500  2013     9    13      601      732            648
## 4        47            500  2013     3     9      547      733            648
## 5        44            500  2013     6     8      544      727            640
## # i 336,771 more rows
## # i 11 more variables: 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>
```

Bài tập 3: Sắp xếp `flights` để tìm chuyến bay có vận tốc nhanh nhất. (Gợi ý: Hãy thử đưa phép tính toán vào bên trong hàm của bạn.)

Ta tính vận tốc (miles/h), chú ý, `air_time` là thời gian bay của máy bay, và đo bằng phút.

```
flights |>
  mutate(speed = distance / (air_time / 60)) |>
  arrange(desc(speed)) |>
  relocate(speed, carrier, origin, dest)
```

```
## # A tibble: 336,776 x 20
##     speed carrier origin dest   year month   day dep_time sched_dep_time dep_delay
##     <dbl> <chr>   <chr>  <chr> <int> <int> <int>    <int>          <int>     <dbl>
## 1 703.385 DL      LGA    ATL    2013     5    25     1709           1700         9
## 2 650.323 EV      EWR    MSP    2013     7     2     1558           1513        45
## 3 648     EV      EWR    GSP    2013     5    13     2040           2025        15
## 4 641.143 EV      EWR    BNA    2013     3    23     1914           1910         4
## 5 591.429 DL      LGA    PBI    2013     1    12     1559           1600        -1
## # i 336,771 more rows
## # i 10 more variables: arr_time <int>, sched_arr_time <int>, arr_delay <dbl>,
## #   flight <int>, tailnum <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

Bài tập 4: Các chuyến bay hàng ngày trong năm 2013, đúng hay không?

Tất cả các ngày trong năm 2013 đều có chuyến bay.

```
flights |>
  distinct(year, month, day) |>
  nrow()
```

```
## [1] 365
```

Chú ý, hàm `nrow()` dùng để truy xuất số dòng có trong một bảng dữ liệu. Tương tự, số cột được truy xuất bằng `ncol()`.

Bài tập 5: Chuyến bay nào có quãng đường xa nhất? Chuyến nào có quãng đường ngắn nhất?

Chuyến bay nào có quãng đường xa nhất

```
flights |>
  arrange(desc(distance)) |>
  relocate(distance)
```

```
## # A tibble: 336,776 x 19
##   distance  year month   day dep_time sched_dep_time dep_delay arr_time
##      <dbl> <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1     4983  2013     1     1      857            900        -3     1516
## 2     4983  2013     1     2      909            900         9     1525
## 3     4983  2013     1     3      914            900        14     1504
```

```
## 4     4983  2013     1     4      900            900         0     1516
## 5     4983  2013     1     5      858            900        -2     1519
## # i 336,771 more rows
## # i 11 more variables: sched_arr_time <int>, arr_delay <dbl>, carrier <chr>,
## #   flight <int>, tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

Chuyến nào có quãng đường ngắn nhất

```
flights |>
  arrange(distance) |>
  relocate(distance)
```

```
## # A tibble: 336,776 x 19
##   distance  year month   day dep_time sched_dep_time dep_delay arr_time
##      <dbl> <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1       17  2013     7    27       NA            106        NA       NA
## 2       80  2013     1     3     2127           2129        -2     2222
## 3       80  2013     1     4     1240           1200        40     1333
## 4       80  2013     1     4     1829           1615       134     1937
## 5       80  2013     1     4     2128           2129        -1     2218
## # i 336,771 more rows
## # i 11 more variables: sched_arr_time <int>, arr_delay <dbl>, carrier <chr>,
## #   flight <int>, tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

Bài tập 6: So sánh các biến `dep_time`, `sched_dep_time` và `dep_delay`, liệu chúng có mối liên hệ nào với nhau.

Thời gian cất cánh thực tế `dep_time` bằng `sched_dep_time` + `dep_delay`.

```
flights |>
  relocate(dep_time, sched_dep_time, dep_delay)
```

```
## # A tibble: 336,776 x 19
##   dep_time sched_dep_time dep_delay  year month   day arr_time sched_arr_time
##      <int>          <int>     <dbl> <int> <int> <int>    <int>          <int>
## 1      517            515         2  2013     1     1      830            819
## 2      533            529         4  2013     1     1      850            830
## 3      542            540         2  2013     1     1      923            850
## 4      544            545        -1  2013     1     1     1004           1022
## 5      554            600        -6  2013     1     1      812            837
## # i 336,771 more rows
## # i 11 more variables: 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>
```

Để kiểm chứng ta có thể làm như sau:

1. Tạo hàm chuyển đổi thời gian hh:mm sang số phút, sau đó **cộng** số phút trễ, sau đó, chuyển đổi kết quả ngược lại dạng hh:mm

```
trans_time_delay <- function(sched, delay){
  temp <- sched %/% 100 * 60 + sched %% 100 + delay
  res <- temp %/% 60 * 100 + temp %% 60
  res <- if_else(res > 2400, res - 2400, res)
  return(res)
}
```

2. Áp dụng hàm vừa viết để tạo một biến mới (`dep_time_guess`), sau đó, so sánh với thời gian cất cánh có sẵn trong dữ liệu (sử dụng hàm `all.equal()`)

```
flights_6 <- flights |>
  mutate(dep_time_guess = trans_time_delay(sched_dep_time, dep_delay)) |>
  relocate(dep_time, dep_time_guess, sched_dep_time, dep_delay)

flights_6
```

```
## # A tibble: 336,776 x 20
##   dep_time dep_time_guess sched_dep_time dep_delay  year month   day arr_time
##      <int>          <dbl>          <int>     <dbl> <int> <int> <int>    <int>
## 1      517            517            515         2  2013     1     1      830
## 2      533            533            529         4  2013     1     1      850
## 3      542            542            540         2  2013     1     1      923
## 4      544            544            545        -1  2013     1     1     1004
## 5      554            554            600        -6  2013     1     1      812
## # i 336,771 more rows
## # i 12 more variables: 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>
```

```
all.equal(flights_6$dep_time, flights_6$dep_time_guess)
```

```
## [1] TRUE
```

Bài tập 7: Tìm hiểu các hàm `starts_with`, `ends_with`, `contains()`. Áp dụng chúng vào trong nhiệm vụ lựa chọn các cột `dep_time`, `dep_delay`, `arr_time` và `arr_delay`.

1. sử dụng `starts_with()`:

```
flights |>
  select(starts_with(c("dep", "arr")))
```

```
## # A tibble: 336,776 x 4
##   dep_time dep_delay arr_time arr_delay
##      <int>     <dbl>    <int>     <dbl>
## 1      517         2      830        11
## 2      533         4      850        20
## 3      542         2      923        33
## 4      544        -1     1004       -18
## 5      554        -6      812       -25
## # i 336,771 more rows
```

2. sử dụng `ends_with()` kết hợp `contains()`:

```
flights |>
  select(ends_with(c("_time", "_delay")), -contains(c("sched", "air")))
```

```
## # A tibble: 336,776 x 4
##   dep_time arr_time dep_delay arr_delay
##      <int>    <int>     <dbl>     <dbl>
## 1      517      830         2        11
## 2      533      850         4        20
## 3      542      923         2        33
## 4      544     1004        -1       -18
## 5      554      812        -6       -25
## # i 336,771 more rows
```

chú ý, ta dùng dấu "-" trước hàm contains() có nghĩa là loại trừ đi.

3. Một cách khác, đó là lấy theo cụm khu vực và dùng contains() để loại bỏ đi các cột không cần thiết:

```
flights |>
  select(dep_time:arr_delay, -contains("sched"))
```

```
## # A tibble: 336,776 x 4
##   dep_time dep_delay arr_time arr_delay
##      <int>     <dbl>    <int>     <dbl>
## 1      517         2      830        11
## 2      533         4      850        20
## 3      542         2      923        33
## 4      544        -1     1004       -18
## 5      554        -6      812       -25
## # i 336,771 more rows
```

Bài tập 8: Hàm any_of() làm gì? Tại sao nó có thể hữu ích khi kết hợp với vectơ này?

```
variables <- c("year", "month", "day", "dep_delay", "arr_delay")
```

Hàm any_of() được dùng để lựa chọn bất kỳ tên biến nào có trong một vector chỉ định, miễn là tên biến đó xuất hiện trong dữ liệu.

```
variables <- c("year", "month", "day", "dep_delay", "arr_delay")
flights |>
  select(any_of(variables))
```

```
## # A tibble: 336,776 x 5
##    year month   day dep_delay arr_delay
##   <int> <int> <int>     <dbl>     <dbl>
## 1  2013     1     1         2        11
## 2  2013     1     1         4        20
## 3  2013     1     1         2        33
## 4  2013     1     1        -1       -18
## 5  2013     1     1        -6       -25
## # i 336,771 more rows
```

Xét một ví dụ khác, trong đó, ta thêm "student" vào trong vector chứ tên cột cần lấy. Rõ ràng, không có cột nào có tên "student" trong bảng dữ liệu, và do đó, không xuất hiện trong kết quả.

```
variables_2 <- c("year", "month", "day", "dep_delay", "arr_delay", "student")
flights |>
  select(any_of(variables_2))
```

```
## # A tibble: 336,776 x 5
##    year month   day dep_delay arr_delay
##   <int> <int> <int>     <dbl>     <dbl>
## 1  2013     1     1         2        11
## 2  2013     1     1         4        20
## 3  2013     1     1         2        33
## 4  2013     1     1        -1       -18
## 5  2013     1     1        -6       -25
## # i 336,771 more rows
```

Bài tập 9: Tại sao đoạn chương trình sau không hoạt động? Lỗi được báo có ý nghĩa gì?

```
flights |> select(tailnum) |>
  arrange(arr_delay)
```

```
## Error in `arrange()`:
## i In argument: `..1 = arr_delay`.
## Caused by error:
## ! object 'arr_delay' not found
```

Đoạn chương trình này không hoạt động. Lỗi được báo có ý nghĩa là biến `arr_delay` không được tìm thấy trong bộ dữ liệu được lựa chọn trước đó. Bởi lẽ hàm `select()` chỉ lựa chọn một biến `tailnum`.

Bài tập 10: Đổi tên cột `air_time` thành `air_time_min` để chỉ rõ đơn vị đo, đồng thời, di chuyển cột này về vị trí bắt đầu của bảng dữ liệu.

```
flights |>
  rename(air_time_min = air_time) |>
  relocate(air_time_min)
```

```
## # A tibble: 336,776 x 19
##   air_time_min  year month   day dep_time sched_dep_time dep_delay arr_time
##          <dbl> <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1          227  2013     1     1      517            515         2      830
## 2          227  2013     1     1      533            529         4      850
## 3          160  2013     1     1      542            540         2      923
## 4          183  2013     1     1      544            545        -1     1004
## 5          116  2013     1     1      554            600        -6      812
## # i 336,771 more rows
## # i 11 more variables: sched_arr_time <int>, arr_delay <dbl>, carrier <chr>,
## #   flight <int>, tailnum <chr>, origin <chr>, dest <chr>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```