R - Practice 07 - v1.1

Sangkon Han(sangkon@pusan.ac.kr)

2023-10-17

Contents

Data Wrangle: dplyr for relational data **1**
Example database 1
Mutating joins 1
inner join 1
left join 2
right join 2
full join 2
Filtering joins 3
semi join 3
anti join 3
Set operations 4
dplyr's additional functions 5

Data Wrangle: dplyr for relational data

Example database

Let's check the database we will be using in this section. `?nycflights13`

Mutating joins

```
## # A tibble: 3 x 2
##     key val
##   <dbl> <chr>
## 1     1 a1
## 2     2 a2
## 3     3 a3
```

```
## # A tibble: 3 x 2
##     key val
##   <dbl> <chr>
## 1     1 b1
## 2     2 b2
## 3     4 b3
```

inner join

```
## # A tibble: 2 x 3
##     key val.x val.y
##   <dbl> <chr> <chr>
## 1     1 a1    b1
## 2     2 a2    b2
```

```
## # A tibble: 16 x 2
##    carrier_name                    n
##    <chr>                       <int>
##  1 United Air Lines Inc.       58665
##  2 JetBlue Airways             54635
##  3 ExpressJet Airlines Inc.    54173
##  4 Delta Air Lines Inc.        48110
##  5 American Airlines Inc.      32729
##  6 Envoy Air                   26397
##  7 US Airways Inc.             20536
##  8 Endeavor Air Inc.           18460
##  9 Southwest Airlines Co.      12275
## 10 Virgin America               5162
## 11 AirTran Airways Corporation  3260
## 12 Alaska Airlines Inc.          714
## 13 Frontier Airlines Inc.        685
## 14 Mesa Airlines Inc.            601
## 15 Hawaiian Airlines Inc.        342
## 16 SkyWest Airlines Inc.          32
```

left join

```
## # A tibble: 3 x 3
##     key val.x val.y
##   <dbl> <chr> <chr>
## 1     1 a1    b1
## 2     2 a2    b2
## 3     3 a3    <NA>
```

right join

```
## # A tibble: 3 x 3
##     key val.x val.y
##   <dbl> <chr> <chr>
## 1     1 a1    b1
## 2     2 a2    b2
## 3     4 <NA>  b3
```

full join

```
## # A tibble: 4 x 3
##     key val.x val.y
##   <dbl> <chr> <chr>
## 1     1 a1    b1
## 2     2 a2    b2
## 3     3 a3    <NA>
## 4     4 <NA>  b3
```

```
## # A tibble: 0 x 3
## # i 3 variables: carrier <chr>, name <chr>, dest <chr>
```

```
## # A tibble: 0 x 3
## # i 3 variables: carrier <chr>, name <chr>, dest <chr>
```

Filtering joins

semi join

```
## # A tibble: 2 x 2
##     key val
##   <dbl> <chr>
## 1     1 a1
## 2     2 a2
```

```
## # A tibble: 3 x 2
##   carrier name
##   <chr>   <chr>
## 1 AA      American Airlines Inc.
## 2 DL      Delta Air Lines Inc.
## 3 VX      Virgin America
```

```
## # A tibble: 86,001 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      542            540         2      923            850
##  2  2013     1     1      554            600        -6      812            837
##  3  2013     1     1      558            600        -2      753            745
##  4  2013     1     1      559            600        -1      941            910
##  5  2013     1     1      602            610        -8      812            820
##  6  2013     1     1      606            610        -4      858            910
##  7  2013     1     1      606            610        -4      837            845
##  8  2013     1     1      615            615         0      833            842
##  9  2013     1     1      623            610        13      920            915
## 10  2013     1     1      628            630        -2     1137           1140
## # i 85,991 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>
```

anti join

```
## # A tibble: 1 x 2
##     key val
##   <dbl> <chr>
## 1     3 a3
```

```
## # A tibble: 0 x 2
## # i 2 variables: carrier <chr>, name <chr>
```

```
## # A tibble: 250,775 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      544            545        -1     1004           1022
##  4  2013     1     1      554            558        -4      740            728
##  5  2013     1     1      555            600        -5      913            854
##  6  2013     1     1      557            600        -3      709            723
##  7  2013     1     1      557            600        -3      838            846
##  8  2013     1     1      558            600        -2      849            851
##  9  2013     1     1      558            600        -2      853            856
## 10  2013     1     1      558            600        -2      924            917
## # i 250,765 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>
```

Set operations

```
## # A tibble: 5 x 4
##   carrier...1 name...2               carrier...3 name...4
##   <chr>       <chr>                  <chr>       <chr>
## 1 9E          Endeavor Air Inc.      AA          American Airlines Inc.
## 2 AS          Alaska Airlines Inc.   B6          JetBlue Airways
## 3 DL          Delta Air Lines Inc.   DL          Delta Air Lines Inc.
## 4 F9          Frontier Airlines Inc. FL          AirTran Airways Corporation
## 5 HA          Hawaiian Airlines Inc. HA          Hawaiian Airlines Inc.
```

```
## # A tibble: 10 x 2
##    carrier name
##    <chr>   <chr>
##  1 9E      Endeavor Air Inc.
##  2 AS      Alaska Airlines Inc.
##  3 DL      Delta Air Lines Inc.
##  4 F9      Frontier Airlines Inc.
##  5 HA      Hawaiian Airlines Inc.
##  6 AA      American Airlines Inc.
##  7 B6      JetBlue Airways
##  8 DL      Delta Air Lines Inc.
##  9 FL      AirTran Airways Corporation
## 10 HA      Hawaiian Airlines Inc.
```

```
## # A tibble: 2 x 2
##   carrier name
##   <chr>   <chr>
## 1 DL      Delta Air Lines Inc.
## 2 HA      Hawaiian Airlines Inc.
```

```
## # A tibble: 3 x 2
##   carrier name
##   <chr>   <chr>
## 1 9E      Endeavor Air Inc.
## 2 AS      Alaska Airlines Inc.
## 3 F9      Frontier Airlines Inc.
```

```
## # A tibble: 8 x 2
##   carrier name
##   <chr>   <chr>
## 1 9E      Endeavor Air Inc.
## 2 AS      Alaska Airlines Inc.
## 3 DL      Delta Air Lines Inc.
## 4 F9      Frontier Airlines Inc.
## 5 HA      Hawaiian Airlines Inc.
## 6 AA      American Airlines Inc.
## 7 B6      JetBlue Airways
## 8 FL      AirTran Airways Corporation
```

dplyr’s additional functions

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 12118
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 25615
```

```
## # A tibble: 1 x 25
##   `fligt id`  year month   day dep_time sched_dep_time dep_delay arr_time
##        <int> <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1        747  2013     1     9      741            745        -4      933
## # i 17 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>,
## #   `origin prev flight` <chr>, `origin test` <lgl>,
## #   `distance successive flights` <dbl>, `distance test` <lgl>,
## #   `distance runing tot` <dbl>
```