Lab-datatable

Lin Pin Tzu (Ruby)

2022-07-19

1) Use and show data.table code to select the variables year, month, day, and hour from the imported flights data

```
flights1 <- fread("nycdata.csv")
flights1
```

```
##         year month day dep_delay arr_delay carrier origin dest air_time
##      1: 2014     1   1        14        13      AA    JFK  LAX      359
##      2: 2014     1   1        -3        13      AA    JFK  LAX      363
##      3: 2014     1   1         2         9      AA    JFK  LAX      351
##      4: 2014     1   1        -8       -26      AA    LGA  PBI      157
##      5: 2014     1   1         2         1      AA    JFK  LAX      350
##     ---
## 253312: 2014    10  31         1       -30      UA    LGA  IAH      201
## 253313: 2014    10  31        -5       -14      UA    EWR  IAH      189
## 253314: 2014    10  31        -8        16      MQ    LGA  RDU       83
## 253315: 2014    10  31        -4        15      MQ    LGA  DTW       75
## 253316: 2014    10  31        -5         1      MQ    LGA  SDF      110
##         distance hour
##      1:     2475    9
##      2:     2475   11
##      3:     2475   19
##      4:     1035    7
##      5:     2475   13
##     ---
## 253312:     1416   14
## 253313:     1400    8
## 253314:      431   11
## 253315:      502   11
## 253316:      659    8
```

```
flights2 <- read_csv("nycdata.csv")
```

```
## Rows: 253316 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): carrier, origin, dest
## dbl (8): year, month, day, dep_delay, arr_delay, air_time, distance, hour
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
flights2
```

```
## # A tibble: 253,316 x 11
##     year month   day dep_delay arr_delay carrier origin dest  air_time distance
##    <dbl> <dbl> <dbl>     <dbl>     <dbl> <chr>   <chr>  <chr>    <dbl>    <dbl>
##  1  2014     1     1        14        13 AA      JFK    LAX        359     2475
##  2  2014     1     1        -3        13 AA      JFK    LAX        363     2475
##  3  2014     1     1         2         9 AA      JFK    LAX        351     2475
##  4  2014     1     1        -8       -26 AA      LGA    PBI        157     1035
##  5  2014     1     1         2         1 AA      JFK    LAX        350     2475
##  6  2014     1     1         4         0 AA      EWR    LAX        339     2454
##  7  2014     1     1        -2       -18 AA      JFK    LAX        338     2475
##  8  2014     1     1        -3       -14 AA      JFK    LAX        356     2475
##  9  2014     1     1        -1       -17 AA      JFK    MIA        161     1089
## 10  2014     1     1        -2       -14 AA      JFK    SEA        349     2422
## # ... with 253,306 more rows, and 1 more variable: hour <dbl>
```

```
flights1[order(year,month,day,hour)]
```

```
##         year month day dep_delay arr_delay carrier origin dest air_time
##      1: 2014     1   1         7        13      B6    JFK  BQN      198
##      2: 2014     1   1        16         8      B6    JFK  PSE      200
##      3: 2014     1   1       110       116      B6    JFK  LAX      349
##      4: 2014     1   1        -7        -6      AA    LGA  ORD      142
##      5: 2014     1   1        -2        -3      AA    LGA  MIA      165
##     ---
## 253312: 2014    10  31       316       289      UA    EWR  ORD      106
## 253313: 2014    10  31        -4         5      B6    JFK  PSE      212
## 253314: 2014    10  31        -3         8      B6    JFK  BQN      202
## 253315: 2014    10  31        -7        -6      B6    JFK  SJU      205
## 253316: 2014    10  31       271       236      UA    LGA  IAH      195
##         distance hour
##      1:     1576    0
##      2:     1617    0
##      3:     2475    0
##      4:      733    5
##      5:     1096    5
##     ---
## 253312:      719   22
## 253313:     1617   23
## 253314:     1576   23
## 253315:     1598   23
## 253316:     1416   23
```

2) Use and show data. table code to produce a table that shows a carrier of DL, an origin of JFK and a destination of SEA

```
flights1[carrier == "DL" & origin == "JFK" & dest == "SEA"]
```

```
##       year month day dep_delay arr_delay carrier origin dest air_time distance
##    1: 2014     1   1        86        79      DL    JFK  SEA      347     2422
##    2: 2014     1   1        -2        -4      DL    JFK  SEA      347     2422
##    3: 2014     1   2         0        11      DL    JFK  SEA      339     2422
```

```
##    4: 2014     1   2        -3         9      DL    JFK  SEA      337     2422
##    5: 2014     1   2        21        19      DL    JFK  SEA      337     2422
##   ---
## 1074: 2014    10  30        -3       -15      DL    JFK  SEA      339     2422
## 1075: 2014    10  31        -6       -26      DL    JFK  SEA      317     2422
## 1076: 2014    10  31        -1        -8      DL    JFK  SEA      338     2422
## 1077: 2014    10  31        -1       -23      DL    JFK  SEA      326     2422
## 1078: 2014    10  31         4       -27      DL    JFK  SEA      318     2422
##       hour
##    1:    9
##    2:   18
##    3:   15
##    4:    7
##    5:   18
##   ---
## 1074:   18
## 1075:    9
## 1076:    6
## 1077:   15
## 1078:   18
```

3) Use and show data.table code to produce a table that shows a carrier of UA, a month of March, and an airtime that is below 330.

```
flights1[carrier == "UA" & month == "3" & air_time < "330"]
```

```
##       year month day dep_delay arr_delay carrier origin dest air_time distance
##    1: 2014     3   1        11        43      UA    EWR  STT      209     1634
##    2: 2014     3   1        47        13      UA    EWR  PBI      133     1023
##    3: 2014     3   1        39        10      UA    EWR  MIA      139     1085
##    4: 2014     3   1        -2       -12      UA    EWR  IAH      197     1400
##    5: 2014     3   1        34        36      UA    EWR  DEN      256     1605
##   ---
## 3434: 2014     3  31         6        -8      UA    EWR  FLL      155     1065
## 3435: 2014     3  31         7        -9      UA    EWR  PBI      135     1023
## 3436: 2014     3  31         1       -21      UA    EWR  RSW      145     1068
## 3437: 2014     3  31         0       -19      UA    EWR  IAH      196     1400
## 3438: 2014     3  31        18        -7      UA    EWR  ORD      108      719
##       hour
##    1:    9
##    2:   19
##    3:   17
##    4:    5
##    5:   16
##   ---
## 3434:   16
## 3435:   10
## 3436:   14
## 3437:   16
## 3438:    6
```

4) Use and show tidyverse code to produce a table that shows a carrier of UA, a month of March, and an airtime that is below 330.

```
flights2 %>%
  select(carrier,month,air_time)%>%
  filter(carrier == "UA", month == "3",air_time < "330") ->ua330
ua330
```

```
## # A tibble: 3,438 x 3
##    carrier month air_time
##    <chr>   <dbl>    <dbl>
##  1 UA          3      209
##  2 UA          3      133
##  3 UA          3      139
##  4 UA          3      197
##  5 UA          3      256
##  6 UA          3      139
##  7 UA          3      123
##  8 UA          3      127
##  9 UA          3      243
## 10 UA          3      140
## # ... with 3,428 more rows
```

5) Use the data.table method to add a variable called speed that is the average air speed of the plane in miles per hour.

```
flights1[, c("speed") := .(distance/(air_time/60))] -> speed
speed
```

6) Use the tidyverse method to add a variable called speed that is the average air speed of the plane in miles per hour.

```
flights2 %>%
  mutate(speed = distance/(air_time/60)) -> mph
mph
```

```
## # A tibble: 253,316 x 12
##     year month   day dep_delay arr_delay carrier origin dest  air_time distance
##    <dbl> <dbl> <dbl>     <dbl>     <dbl> <chr>   <chr>  <chr>    <dbl>    <dbl>
##  1  2014     1     1        14        13 AA      JFK    LAX        359     2475
##  2  2014     1     1        -3        13 AA      JFK    LAX        363     2475
##  3  2014     1     1         2         9 AA      JFK    LAX        351     2475
##  4  2014     1     1        -8       -26 AA      LGA    PBI        157     1035
##  5  2014     1     1         2         1 AA      JFK    LAX        350     2475
##  6  2014     1     1         4         0 AA      EWR    LAX        339     2454
##  7  2014     1     1        -2       -18 AA      JFK    LAX        338     2475
##  8  2014     1     1        -3       -14 AA      JFK    LAX        356     2475
##  9  2014     1     1        -1       -17 AA      JFK    MIA        161     1089
## 10  2014     1     1        -2       -14 AA      JFK    SEA        349     2422
## # ... with 253,306 more rows, and 2 more variables: hour <dbl>, speed <dbl>
```

7) Show and use coding to change the carrier abbreviation of UA to UniitedAir

7a) data.table method

```
flights1[carrier == "UA", carrier := "UniitedAir"]
flights1
```

```
##         year month day dep_delay arr_delay    carrier origin dest air_time
##      1: 2014     1   1        14        13         AA    JFK  LAX      359
##      2: 2014     1   1        -3        13         AA    JFK  LAX      363
##      3: 2014     1   1         2         9         AA    JFK  LAX      351
##      4: 2014     1   1        -8       -26         AA    LGA  PBI      157
##      5: 2014     1   1         2         1         AA    JFK  LAX      350
##     ---
## 253312: 2014    10  31         1       -30 UniitedAir    LGA  IAH      201
## 253313: 2014    10  31        -5       -14 UniitedAir    EWR  IAH      189
## 253314: 2014    10  31        -8        16         MQ    LGA  RDU       83
## 253315: 2014    10  31        -4        15         MQ    LGA  DTW       75
## 253316: 2014    10  31        -5         1         MQ    LGA  SDF      110
##         distance hour    speed
##      1:     2475    9 413.6490
##      2:     2475   11 409.0909
##      3:     2475   19 423.0769
##      4:     1035    7 395.5414
##      5:     2475   13 424.2857
##     ---
## 253312:     1416   14 422.6866
## 253313:     1400    8 444.4444
## 253314:      431   11 311.5663
## 253315:      502   11 401.6000
## 253316:      659    8 359.4545
```

7b tidyverse method (Use a sequence of dplyr commands so that you can see the change in your table)

```
flights2 %>%
  mutate(carrier = recode(carrier, "UA" = "UniitedAir")) %>%
   filter(carrier=="UniitedAir")
```

```
## # A tibble: 46,267 x 11
##     year month   day dep_delay arr_delay carrier  origin dest  air_time distance
##    <dbl> <dbl> <dbl>     <dbl>     <dbl> <chr>    <chr>  <chr>    <dbl>    <dbl>
##  1  2014     1     1         9        -2 Uniited~ EWR    HNL        630     4963
##  2  2014     1     1        25        17 Uniited~ EWR    TPA        149      997
##  3  2014     1     1        49        57 Uniited~ EWR    TPA        157      997
##  4  2014     1     1         0         9 Uniited~ EWR    TPA        171      997
##  5  2014     1     1         8        -1 Uniited~ EWR    SAT        235     1569
##  6  2014     1     1        43        42 Uniited~ EWR    MIA        155     1085
##  7  2014     1     1        10         4 Uniited~ EWR    PBI        155     1023
##  8  2014     1     1         0        11 Uniited~ EWR    TPA        162      997
##  9  2014     1     1         5        -3 Uniited~ EWR    RSW        165     1068
## 10  2014     1     1         0         8 Uniited~ LGA    ORD        131      733
```

```
## # ... with 46,257 more rows, and 1 more variable: hour <dbl>
```