Tarea - Filtrando los Datos con dplyr

Jesus Mudarra Luján

2022-10-30

Pregunta 1

Encuentra todos los vuelos que llegaron más de una hora tarde de lo previsto.

```
filter(flights, arr_delay>60)
```

```
## # A tibble: 27,789 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      811        630     101    1047     830     137 MQ
##  2  2013     1     1      848       1835     853    1001    1950     851 MQ
##  3  2013     1     1      957        733     144    1056     853     123 UA
##  4  2013     1     1     1114        900     134    1447    1222     145 UA
##  5  2013     1     1     1120        944      96    1331    1213      78 EV
##  6  2013     1     1     1255       1200      55    1451    1330      81 MQ
##  7  2013     1     1     1301       1150      71    1518    1345      93 MQ
##  8  2013     1     1     1337       1220      77    1649    1531      78 B6
##  9  2013     1     1     1342       1320      22    1617    1504      73 EV
## 10  2013     1     1     1400       1250      70    1645    1502     103 EV
## # ... with 27,779 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Pregunta 2

Encuentra todos los vuelos que volaron hacia San Francisco (aeropuertos SFO y OAK)

```
filter(flights, dest=="SFO"|dest=="OAK")
```

```
## # A tibble: 13,643 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      558        600      -2     923     937     -14 UA
##  2  2013     1     1      611        600      11     945     931      14 UA
##  3  2013     1     1      655        700      -5    1037    1045      -8 DL
##  4  2013     1     1      729        730      -1    1049    1115     -26 VX
##  5  2013     1     1      734        737      -3    1047    1113     -26 B6
##  6  2013     1     1      745        745       0    1135    1125      10 AA
##  7  2013     1     1      746        746       0    1119    1129     -10 UA
##  8  2013     1     1      803        800       3    1132    1144     -12 UA
##  9  2013     1     1      826        817       9    1145    1158     -13 UA
## 10  2013     1     1     1029       1030      -1    1427    1355      32 AA
```

```
## # ... with 13,633 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Pregunta 3

Encuentra todos los vuelos operados por United American (UA) o por American Airlines (AA)

```
filter(flights, carrier=="UA"|carrier=="AA")
```

```
## # A tibble: 91,394 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      517        515       2     830     819      11 UA
##  2  2013     1     1      533        529       4     850     830      20 UA
##  3  2013     1     1      542        540       2     923     850      33 AA
##  4  2013     1     1      554        558      -4     740     728      12 UA
##  5  2013     1     1      558        600      -2     753     745       8 AA
##  6  2013     1     1      558        600      -2     924     917       7 UA
##  7  2013     1     1      558        600      -2     923     937     -14 UA
##  8  2013     1     1      559        600      -1     941     910      31 AA
##  9  2013     1     1      559        600      -1     854     902      -8 UA
## 10  2013     1     1      606        610      -4     858     910     -12 AA
## # ... with 91,384 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Pregunta 4

Encuentra todos los vuelos que salieron los meses de primavera (Abril, Mayo y Junio)

```
filter(flights, month %in% c(4,5,6))
```

```
## # A tibble: 85,369 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     4     1      454        500      -6     636     640      -4 US
##  2  2013     4     1      509        515      -6     743     814     -31 UA
##  3  2013     4     1      526        530      -4     812     827     -15 UA
##  4  2013     4     1      534        540      -6     833     850     -17 AA
##  5  2013     4     1      542        545      -3     914     920      -6 B6
##  6  2013     4     1      543        545      -2     921     927      -6 B6
##  7  2013     4     1      551        600      -9     748     659      49 US
##  8  2013     4     1      552        600      -8     641     701     -20 US
##  9  2013     4     1      553        600      -7     725     735     -10 MQ
## 10  2013     4     1      554        600      -6     752     805     -13 EV
## # ... with 85,359 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Pregunta 5

Encuentra todos los vuelos que llegaron más de una hora tarde, pero salieron con menos de una hora de retraso.

```
filter(flights, dep_delay<60&arr_delay>60)
```

```
## # A tibble: 4,956 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1     1255       1200      55    1451    1330      81 MQ
##  2  2013     1     1     1342       1320      22    1617    1504      73 EV
##  3  2013     1     1     1402       1323      39    1650    1526      84 EV
##  4  2013     1     1     1411       1315      56    1717    1611      66 B6
##  5  2013     1     1     1424       1349      35    1701    1556      65 EV
##  6  2013     1     1     1428       1329      59    1803    1640      83 B6
##  7  2013     1     1     1558       1534      24    1808    1703      65 EV
##  8  2013     1     1     1604       1510      54    1817    1710      67 MQ
##  9  2013     1     1     1608       1535      33    2002    1850      72 AA
## 10  2013     1     1     1630       1548      42    1902    1755      67 EV
## # ... with 4,946 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Pregunta 6

Encuentra todos los vuelos que salieron con más de una hora de retraso, pero consiguieron llegar con menos de 30 minutos de retraso (el avión aceleró en el aire)

```
filter(flights, dep_delay>60&arr_delay<30)
```

```
## # A tibble: 181 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     3     1850       1745      65    2148    2120      28 AA
##  2  2013     1     3     1950       1845      65    2228    2227       1 B6
##  3  2013     1     6     1019        900      79    1558    1530      28 HA
##  4  2013     1     7     1543       1430      73    1758    1735      23 AA
##  5  2013     1    12     1706       1600      66    1949    1927      22 DL
##  6  2013     1    12     1953       1845      68    2154    2137      17 9E
##  7  2013     1    19     1456       1355      61    1636    1615      21 EV
##  8  2013     1    21     1531       1430      61    1843    1815      28 DL
##  9  2013     1    21     1648       1545      63    1939    1910      29 AA
## 10  2013    10    10     1938       1835      63    2158    2148      10 AS
## # ... with 171 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Pregunta 7

Encuentra todos los vuelos que salen entre medianoche y las 7 de la mañana (vuelos nocturnos).

```
filter(flights, hour>=0, hour<7)
```

```
## # A tibble: 27,905 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      517        515       2     830     819      11 UA
##  2  2013     1     1      533        529       4     850     830      20 UA
##  3  2013     1     1      542        540       2     923     850      33 AA
##  4  2013     1     1      544        545      -1    1004    1022     -18 B6
##  5  2013     1     1      554        600      -6     812     837     -25 DL
##  6  2013     1     1      554        558      -4     740     728      12 UA
##  7  2013     1     1      555        600      -5     913     854      19 B6
##  8  2013     1     1      557        600      -3     709     723     -14 EV
##  9  2013     1     1      557        600      -3     838     846      -8 B6
## 10  2013     1     1      558        600      -2     753     745       8 AA
## # ... with 27,895 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Pregunta 8

Investiga el uso de la función `between()` de `dplyr`. ¿Qué hace? ¿Puedes usarlo para resolver la sintaxis necesaria para responder alguna de las preguntas anteriores?

Se trata de un atajo para $x \geq$ izquierda y $x \leq$ derecha, implementado eficientemente en C++ para valores locales, y traducido al SQL apropiado para tablas remotas.

Ejemplo resolviendo el ejercicio 4:

```
filter(flights, between(month,4,6))
```

```
## # A tibble: 85,369 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     4     1      454        500      -6     636     640      -4 US
##  2  2013     4     1      509        515      -6     743     814     -31 UA
##  3  2013     4     1      526        530      -4     812     827     -15 UA
##  4  2013     4     1      534        540      -6     833     850     -17 AA
##  5  2013     4     1      542        545      -3     914     920      -6 B6
##  6  2013     4     1      543        545      -2     921     927      -6 B6
##  7  2013     4     1      551        600      -9     748     659      49 US
##  8  2013     4     1      552        600      -8     641     701     -20 US
##  9  2013     4     1      553        600      -7     725     735     -10 MQ
## 10  2013     4     1      554        600      -6     752     805     -13 EV
## # ... with 85,359 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Pregunta 9

¿Cuántos vuelos tienen un valor desconocido de dep_time?

```r
sumNA <- sum(is.na(flights$dep_time))
percNA <- sumNA/length(rownames(flights))*100
```

En dep_time hay 8255 valores desconocidos, lo que representa un 2.4511842%.

Pregunta 10

¿Qué variables del dataset contienen valores desconocidos? ¿Qué representan esas filas donde faltan los datos?

Estas son las variables que contienen valores desconocidos.

```r
apply(flights, MARGIN = 2, FUN = function(x) sum(is.na(x)))
```

```
##           year          month            day       dep_time sched_dep_time
##              0              0              0           8255              0
##      dep_delay       arr_time sched_arr_time      arr_delay        carrier
##           8255           8713              0           9430              0
##         flight        tailnum         origin           dest       air_time
##              0           2512              0              0           9430
##       distance           hour         minute      time_hour
##              0              0              0              0
```

Pregunta 11

Ahora vas a sorprenderte con la magia oscura. . . Contesta qué dan las siguientes condiciones booleanas

```r
NA*0
```

```
## [1] NA
```

```r
NA|TRUE
```

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

```r
FALSE&NA
```

```
## [1] FALSE
```

Intenta establecer la regla general para saber cuando es o no es `NA` (cuidado con `NA*0`)

Siempre se va a obtener un valor `NA` con cualquier tipo de operación excepto cuando se realice la operación booleana `OR` donde se obtendrá un valor de `TRUE` al dar por válido el valor no `NA`.