Tarea 2

Representación y tabulación de datos

Julio David Ruiz Mendoza

Mayo 2021

Enunciado

Vamos a responder preguntas sobre los datos que se encuentran en el fichero `vuelos.csv` que contiene datos sobre horarios, retrasos, aviones, tiempos y distancias de vuelos que salen de Houston los primeros meses de 2011.

Es obligatorio utilizar las funciones del paquete `dplyr` y recomendable utilizar pipes `%>%`.

1. Carga la librería `dplyr`

```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

2. Descarga el fichero de datos `vuelos.csv` situado en http://gauss.inf.um.es/datos/ en local o localiza la url donde se encuentra.

Lee el fichero (desde local o desde la url) usando correctamente los argumentos de la función `read.table()`: `header`, `sep`, `dec`. Guarda los datos en una variable llamada `vuelos`.

```
vuelos <- read.table(file = "vuelos.csv", header = TRUE, sep = ',',
                     dec = '.', stringsAsFactors = TRUE)
```

3. Selecciona los vuelos con destino `SFO` u `OAK` utilizando las funciones del paquete dplyr. ¿Con cuántos vuelos nos quedamos?

```
head(filter(vuelos,dest == "SFO" | dest == "OAK"))
```

```
##           date hour minute  dep  arr dep_delay arr_delay carrier flight dest
## 373 2011-01-31    8     51  851 1052         1       -27      CO    170  SFO
## 389 2011-01-31   11     29 1129 1351         4         1      CO    270  SFO
## 402 2011-01-31   14     32 1432 1656         7         5      CO    370  SFO
## 436 2011-01-31   17     48 1748 2001         3        -4      CO    570  SFO
## 467 2011-01-31   21     43 2143 2338        50        24      CO    770  SFO
## 468 2011-01-31    7     29  729 1002        -1         2      CO    771  SFO
##      plane cancelled time dist
## 373 N35407         0  225 1635
```

```
## 389 N37420         0  228 1635
## 402 N27213         0  229 1635
## 436 N75436         0  236 1635
## 467 N37281         0  224 1635
## 468 N26226         0  237 1635
```

```
nrow(filter(vuelos,dest == "SFO" | dest == "OAK"))
```

```
## [1] 1121
```

Hay 1121 vuelos con esos destinos.

4. Selecciona los vuelos que se han retrasado más de una hora. ¿Cuál es el destino que más se retrasa en proporción al número de vuelos?

```
vuelos_del <- filter(vuelos, dep_delay > 60)
a <- dplyr::count(vuelos_del,dest)
a[which.max(a$n),]
```

```
##   dest   n
## 5  ATL 163
```

```
vuelos_del <- filter(vuelos, arr_delay > 60)
b <- dplyr::count(vuelos_del,dest)
b[which.max(b$n),]
```

```
##    dest   n
## 77  ORD 170
```

El destino que más se retrasa en salida es ATL con 163 retrasos, a la llegada ORD con 170.

5. Encuentra 4 maneras diferentes de utilizar la función `select` para seleccionar las variables relacionadas con los retrasos (delay)

```
head(  select(vuelos, dep_delay, arr_delay)  )
```

```
##   dep_delay arr_delay
## 1         0       -10
## 2         1        -9
## 3        -8        -8
## 4         3         3
## 5         5        -3
## 6        -1        -7
```

```
head(  select(vuelos, ends_with('delay'))  )
```

```
##   dep_delay arr_delay
## 1         0       -10
## 2         1        -9
## 3        -8        -8
## 4         3         3
## 5         5        -3
## 6        -1        -7
```

```
head(  select(vuelos, matches('delay'))  )
```

```
##   dep_delay arr_delay
## 1         0       -10
## 2         1        -9
## 3        -8        -8
```

```
## 4         3         3
## 5         5        -3
## 6        -1        -7
```

```
head(  select(vuelos, contains('delay'))  )
```

```
##   dep_delay arr_delay
## 1         0       -10
## 2         1        -9
## 3        -8        -8
## 4         3         3
## 5         5        -3
## 6        -1        -7
```

6. Agrupa los vuelos por fecha y calcula: media, mediana y cuartil 75 de los retrasos en los vuelos por hora.

```
summarise(group_by(vuelos,date,hour), media = mean(dep_delay),
          mediana=median(dep_delay),
          cuartil_75=quantile(dep_delay ,0.75, na.rm = TRUE), n=n())
```

```
## `summarise()` has grouped output by 'date'. You can override using the `.groups` argument.
```

```
## # A tibble: 2,362 x 6
## # Groups:   date [120]
##    date        hour media mediana cuartil_75     n
##    <fct>      <int> <dbl>   <dbl>      <dbl> <int>
##  1 2011-01-01     0  2        2         2        1
##  2 2011-01-01     5 -1.75    -1.5      -1        4
##  3 2011-01-01     6 -1.86     0         0.5      7
##  4 2011-01-01     7  5.23     0.5       3       30
##  5 2011-01-01     8  4.22     0.5       5.5     18
##  6 2011-01-01     9  4.10     2         7       29
##  7 2011-01-01    10  2.43     0         2       47
##  8 2011-01-01    11 10.6      3         8.5     47
##  9 2011-01-01    12  3.86     0.5       7.25    44
## 10 2011-01-01    13  4.78     2         9       32
## # ... with 2,352 more rows
```

7. Utilizando pipes calcula la media de retraso en los vuelos por día y hora, la cantidad de vuelos por día y hora y luego muestra solo los casos para los cuales haya más de 10.

```
a <- vuelos %>%
  group_by(date,hour) %>%
  summarise(media =  mean(dep_delay), n= n())
```

```
## `summarise()` has grouped output by 'date'. You can override using the `.groups` argument.
```

```
a
```

```
## # A tibble: 2,362 x 4
## # Groups:   date [120]
##    date        hour media     n
##    <fct>      <int> <dbl> <int>
##  1 2011-01-01     0  2        1
##  2 2011-01-01     5 -1.75     4
##  3 2011-01-01     6 -1.86     7
##  4 2011-01-01     7  5.23    30
##  5 2011-01-01     8  4.22    18
```

```
##  6 2011-01-01     9  4.10    29
##  7 2011-01-01    10  2.43    47
##  8 2011-01-01    11 10.6     47
##  9 2011-01-01    12  3.86    44
## 10 2011-01-01    13  4.78    32
## # ... with 2,352 more rows
```

```
a %>%
  filter(n>10)
```

```
## # A tibble: 1,905 x 4
## # Groups:   date [120]
##    date        hour media     n
##    <fct>      <int> <dbl> <int>
##  1 2011-01-01     7  5.23    30
##  2 2011-01-01     8  4.22    18
##  3 2011-01-01     9  4.10    29
##  4 2011-01-01    10  2.43    47
##  5 2011-01-01    11 10.6     47
##  6 2011-01-01    12  3.86    44
##  7 2011-01-01    13  4.78    32
##  8 2011-01-01    14  9.58    43
##  9 2011-01-01    15  9.72    40
## 10 2011-01-01    16 15.5     27
## # ... with 1,895 more rows
```