EXERCICI – SCRIPTING PER BASES DE DADES

1. Inici
   - **a.** Instal·la el paquet "nycflights13"
   - **b.** Carrega el paquet "nycflights13".

   Es tracta d'una paquet amb dades sobre els vols que van volar des dels aeroports JFK, LGA i EWR de Nova York l'any 2013.

   - **c.** Busca informació sobre el paquet "nycflights13" a Internet: quantes i quines bases de dades conté?

     Ens interessen les variables:
     - "hour", "minute": l'hora i minut de la sortida del vol,
     - "arr_delay": el retard del vol,
     - " dest": destinació del vol.
   - **d.** Carrega la base de dades adequada (la que conté les variables d'interès), assigna-la a "vols" i investiga-la una mica [str(), head(), summary()]

2. Retard vs. Temps

La idea general és trobar una relació entre el retard dels vols i l'hora de sortida dels vols, per poder descobrir si és millor volar al matí, a la tarda o a la nit per evitar retards.

   - **a.** Crea una nova variable "time" dins la base de dades "vols", que contingui l'hora i minut com un sol valor decimal (de l'estil 1.01, 1.10, 1.50 hores).
   - **b.** Calcula el retard mitjà i el nombre de vols per hora per a cada valor de "time" diferent. El resultat hauria de ser una dataframe anomenada "retard.per.hora" de 3 columnes, "time", "retard" i "n", amb tantes files com valors de "time" diferents existeixin. Utilitzar el paquet "dplyr" facilita les coses!
   - **c.** Visualitza el resultat amb l'ordre "View(retard.per.hora)".
   - **d.** BONUS! Hi retornarem després de practicar amb ggplot2, però es pot provar igualment. Grafica el retard mitjà versus el temps. Escala la mida dels punts segons el nombre de vols. Quines conclusions se'n poden treure?

3. Retard vs. Temps

a. De manera similar a com s'ha construit la dataframe "retard.per.hora" construirem una dataframe anomenada "retard.per.dest". El resultat serà una dataframe amb 3 columnes, "dest", "retard" i "n" que contenen, respectivament, el nom de la destinació del vol, el retard del vol i el nombre de vols que han volat a cada destinació. La dataframe tindrà tantes files com destinacions diferents hi hagi.

b. Visualitza "retard.per.dest".

c. Uneix-li la informació d'aeroports (noms i llocs en forma de latitud i longitud). Es pot fer mitjançant la comanda "left_join" del paquet "dplyr" o bé amb la comanda "merge" de R base. Cal tenir en compte que en una dataframe els noms dels aeroports de destinació estan a la variable "dest" i en l'altra a la variable "faa".

```
retard.per.dest <-left_join(x = retard.per.dest, y = ...,
        by = c("dest" = "faa"))
```

d. BONUS! Grafica latitud versus longitud i escala la mida dels punts segons el nombre de vols.

```
ggplot(data = ..., aes(..., ..., size = n)) + geom_point()
```

e. BONUS! Afegeix un mapa dels Estats Units sobre el gràfic. Instal·la i carrega el paquet "maps". Afegeix "+ borders(database = "state", size = 0.5) + geom_point()" a la crida del gràfic.