Ejercicio-ggplot2-dplyr

Felipe Ortega, Isaac Martín

7 de octubre de 2016

Introducción

El paquete `nycflights13`, disponible en CRAN, contiene datos sobre 336.776 vuelos que despegaron de alguno de los tres aeropuertos que dan servicio a la ciudad de Nueva York (EE.UU.) en 2013, procedentes del Bureau of Transport Statistics:

- Aeropuerto Internacional Libertad de Newark (EWR).
- Aeropuerto Internacional John. F. Kennedy (JFK).
- Aeropuerto Internacional de La Guardia (LGA).

El conjunto principal de datos sobre los vuelos está disponible en el `data.frame flights`, dentro de este paquete. Adicionalmente, su autor (Hadley Wickham) también ha incluido datos sobre los propios aeropuertos, condiciones meteorológicas, etc. Para más detalles, ver archivo de descripción del paquete con el comando `?nycflights13`.

Preparación del ejercicio

Durante el ejercicio, se utilizarán las bibliotecas `ggplot2` y `dplyr`, ya introducidas en clase.

Nota importante 1: Se recomienda revisar y practicar con los ejemplos del documento de introducción a `dplyr` antes de realizar este ejercicio, así como los ejemplos incluidos en el seminario de H. Wickham sobre "Tidy Data", enlazado en la sección referencias del Tema 2 en Aula Virtual.

Nota importante 2: intente utilizar el operador %>% (*forward pipe*) para el código de resolución de todos los ejercicios.

```
# Importamos bibliotecas y datos
library(ggplot2)
library(dplyr)
library(nycflights13)
```

Ejercicio 1 (*30 puntos*)

Utiliza las funciones incluidas en el paquete `dplyr`, para responder a las siguientes preguntas:

a) ¿Cuantos vuelos se realizan en total cada mes?
b) ¿Qué aeropuerto acumula el mayor número de salidas de vuelos en todo el año?
c) ¿Qué compañía acumula el mayor número de salida de vuelos en los meses de verano (jun-sep.)?
d) ¿Qué compañia acumula más tiempo de vuelo en todo el año?
e) ¿Qué compañía registra los mayores retrasos de salida de sus vuelos? ¿Tienen los retrasos alguna correlación con la duración de los vuelos?

```
## Cuantos vuelos se realizan cada mes
flights %>%
  group_by(month) %>%
  summarise(vuelos = n())
```

```
## # A tibble: 12 × 2
##    month vuelos
```

```
##    <int>  <int>
## 1      1  27004
## 2      2  24951
## 3      3  28834
## 4      4  28330
## 5      5  28796
## 6      6  28243
## 7      7  29425
## 8      8  29327
## 9      9  27574
## 10    10  28889
## 11    11  27268
## 12    12  28135
```

```
## Que aeropuerto tiene el mayor numero de salidas
flights %>%
  group_by(origin) %>%
  summarise(salidas = n()) %>%
  top_n(1)
```

```
## Selecting by salidas
```

```
## # A tibble: 1 × 2
##   origin salidas
##    <chr>   <int>
## 1    EWR  120835
```

```
## Compa?ia con mayor numero de salidas en verano
flights %>%
  filter(month==c(6,7,8,9)) %>%
  group_by(carrier) %>%
  summarise(verano = n()) %>%
  top_n(1)
```

```
## Selecting by verano
```

```
## # A tibble: 1 × 2
##   carrier verano
##     <chr>  <int>
## 1      UA   4848
```

```
## Compa?ia con mayor tiempo de vuelo en todo el a?o
flights %>%
  group_by(carrier) %>%
  summarise(tiempo = sum(air_time,na.rm=TRUE)) %>%
  top_n(1)
```

```
## Selecting by tiempo
```

```
## # A tibble: 1 × 2
##   carrier   tiempo
##     <chr>    <dbl>
## 1      UA 12237728
```

```
## Compa?ia con mayor retraso en la salida de los vuelos
flights %>%
  group_by(carrier) %>%
  filter(dep_delay>0) %>%
```

```
  summarise(tiempo = sum(dep_delay,na.rm=TRUE)) %>%
  top_n(1)
```

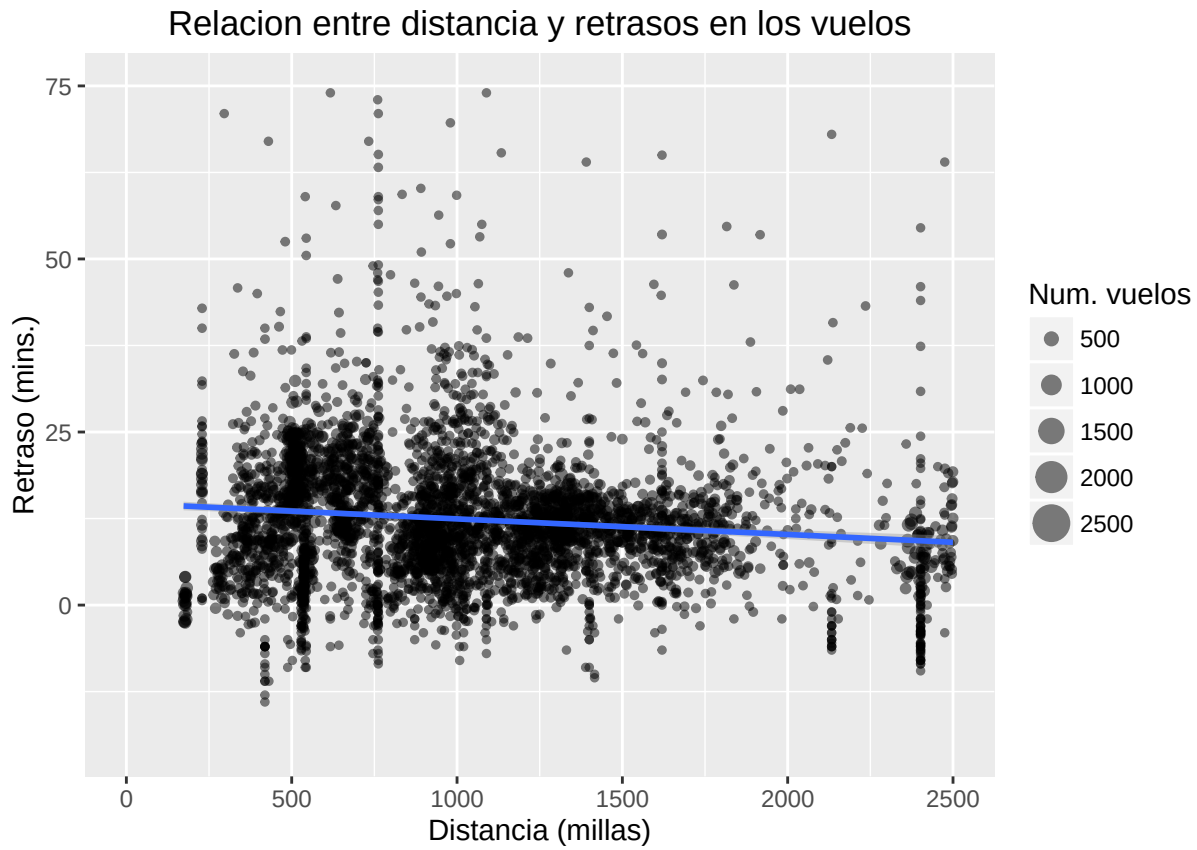
```
## Selecting by tiempo
```

```
## # A tibble: 1 × 2
##   carrier  tiempo
##     <chr>   <dbl>
## 1      EV 1164581
```

```
## Tiene alguna relacion los retrasos en las salidas con la distancia de los vuelos
by_tailnum <- group_by(flights, tailnum)
delay <- summarise(by_tailnum,
                   count = n(),
                   dist = mean(distance, na.rm = TRUE),
                   delay = mean(dep_delay, na.rm = TRUE))

ggplot(delay, aes(dist, delay)) +
  geom_point(aes(size = count), alpha = 1/2) +
  labs(x="Distancia (millas)", y="Retraso (mins.)") +
  xlim(c(0,2500)) +
  ylim(c(-20,75)) +
  geom_smooth(method = 'gam') +
  scale_size_area() +
  ggtitle("Relacion entre distancia y retrasos en los vuelos") +
  scale_radius(name="Num. vuelos")
```

```
## Scale for 'size' is already present. Adding another scale for 'size',
## which will replace the existing scale.
```



Los datos faltantes los he omitido a la hora de hacer la suma en los dos ultimos apartados puesto que no se puede saber la causa por la que no se tienen esos datos. En ambos casos el numero de datos faltantes oscilan sobre los 9000 que supone aproximadamente un cuarto de los datos totales pero que aun dejan una gran cantidad de datos para hacer el analisis fiable en cierto modo.

En el ultimo apartado para mirar la compa?ia con mayor retraso en la salida de los vuelos solo he tenido en cuenta los retrasos, no los que hayan salido antes de tiempo, suponiendo que el objetivo de toda empresa siempre es tener el menor retraso posible en sus vuelos. En cuanto a la relacion entre los retrasos y la distancia de los vuelos se puede apreciar una ligerisima correlacion negativa, apenas inexistente.

Ejercicio 2 (*30 puntos*)

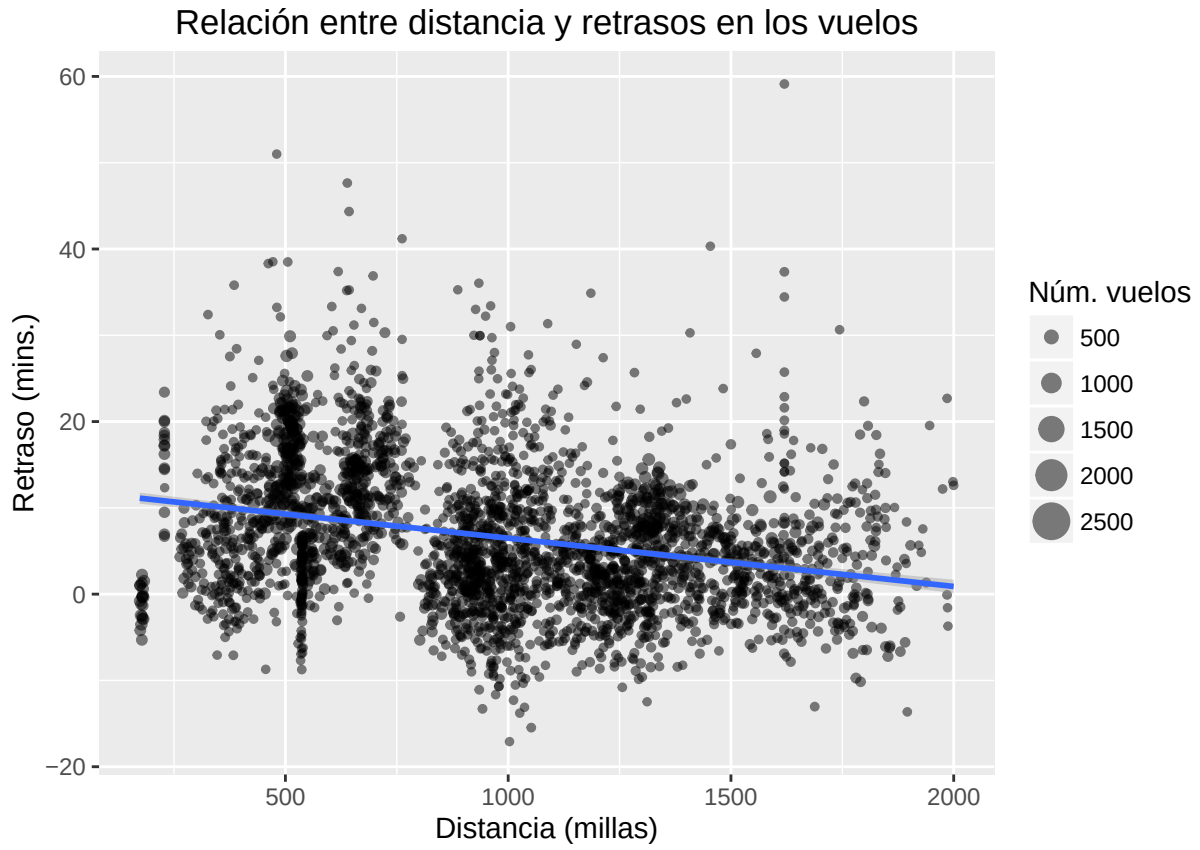
La siguiente figura, tomada de la introducción a dplyr, muestra un gráfico en `ggplot2` de la relación entre distancia de los vuelos y retraso experimentado para todos los aeropuertos de NYC.

```
by_tailnum <- group_by(flights, tailnum)
delay <- summarise(by_tailnum,
  count = n(),
  dist = mean(distance, na.rm = TRUE),
  delay = mean(arr_delay, na.rm = TRUE))
delay <- filter(delay, count > 20, dist < 2000)

# Interestingly, the average delay is only slightly related to the
# average distance flown by a plane.
ggplot(delay, aes(dist, delay)) +
  geom_point(aes(size = count), alpha = 1/2) +
  labs(x="Distancia (millas)", y="Retraso (mins.)") +
```

```
  geom_smooth(method = 'gam') +
  scale_size_area() +
  ggtitle("Relación entre distancia y retrasos en los vuelos") +
  scale_radius(name="Núm. vuelos")
```

```
## Scale for 'size' is already present. Adding another scale for 'size',
## which will replace the existing scale.
```



A la vista del resultado, parece que exista una cierta correlación negativa, aunque no muy fuerte, entre ambas variables. Sin embargo, veamos que sucede si desglosamos los datos utilizando otras variables disponibles.

En este ejercicio, se propone **representar el retraso de llegadas en función de la distancia recorrida**, utilizando una gráfica como la anterior, pero desglosado por meses (es decir, una gráfica como la anterior para cada mes).

La solución óptima debería construir un panel de 12 gráficas, una para cada mes. Cada gráfica se debe etiquetar con el nombre abreviado de ese mes, no con el número de mes. Además, se debe presentar las gráficas en el orden correcto de los meses del calendario (primero el gráfico de enero, luego febrero, etc.), no por orden alfabético de los nombres del mes.

¿Qué conclusiones puedes extraer a la vista de estos gráficos? Intenta ofrecer argumentos basados en los resultados obtenidos para elaborar la respuesta.

```
by_tailnum <- flights %>%
  filter(arr_delay>0) %>%
  mutate(mes = month.abb[month]) %>%
  group_by(tailnum, mes)
delay <- summarise(by_tailnum,
```

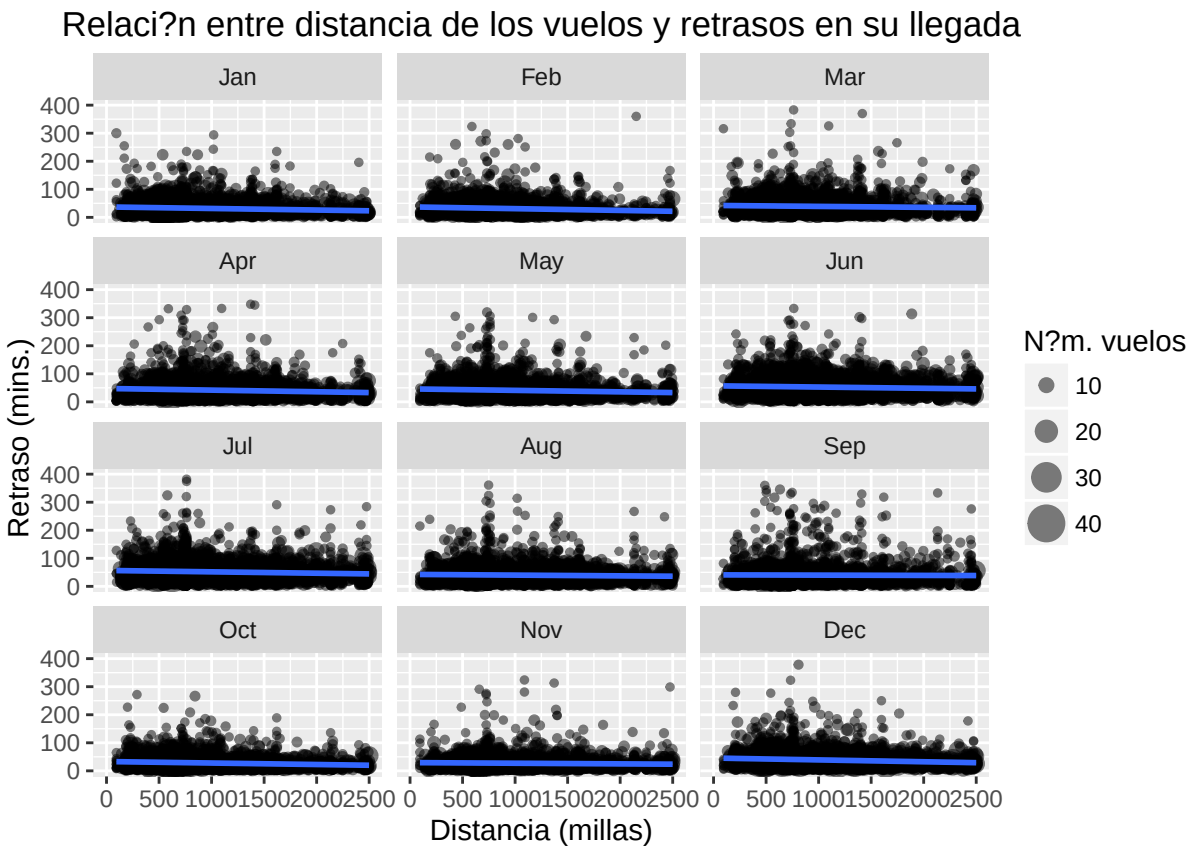
```
                    count = n(),
                    dist = mean(distance, na.rm = TRUE),
                    delay = mean(arr_delay, na.rm = TRUE))
delay$mes<- factor(delay$mes,month.abb)

ggplot(delay, aes(dist, delay)) +
  geom_point(aes(size = count), alpha = 1/2) +
  xlim(c(0,2500)) +
  ylim(c(0,400)) +
  labs(title="Relaci?n entre distancia de los vuelos y retrasos en su llegada", x="Distancia (millas)",
  geom_smooth(method = 'gam') +
  scale_size_area() +
  scale_radius(name="N?m. vuelos") +
  facet_wrap(~mes, ncol = 3)
```

```
## Scale for 'size' is already present. Adding another scale for 'size',
## which will replace the existing scale.
```

```
## Warning: Removed 1023 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1023 rows containing missing values (geom_point).
```



En vista del grafico obtenido no se observa ninguna relacion clara para ninguno de los meses entre la distancia del vuelo y el retraso en la salida. Al igual que antes puede verse una peque?a correlacion negativa en todos los meses pero desde mi punto de vista totalmente insignificante.

Ejercicio 3 (*20 puntos*)

Representar el retrasos de salida de los vuelos que parten del aeropuerto JFK (código 'JFK'), desglosado por meses (como en el ejercicio anterior). Se mostrarán solo los vuelos domésticos, imponiendo como condición de filtrado de datos: `distancia recorrida < 1.000 millas`.

¿Qué conclusiones puedes extraer a la vista de estos gráficos?

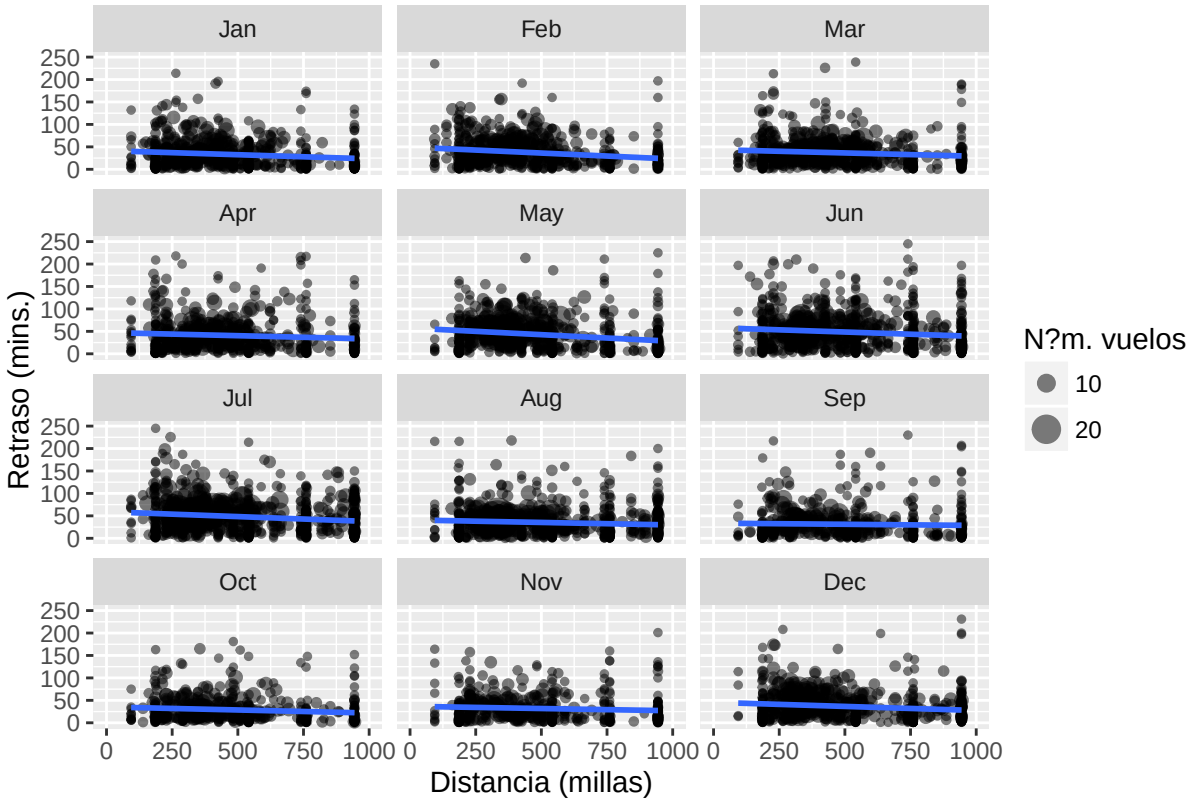
```
by_tailnum <- flights %>%
  filter(dep_delay>0, distance<1000, origin == 'JFK') %>%
  mutate(mes = month.abb[month]) %>%
  group_by(tailnum, mes)
delay <- summarise(by_tailnum,
                   count = n(),
                   dist = mean(distance, na.rm = TRUE),
                   delay = mean(dep_delay, na.rm = TRUE))
delay$mes<- factor(delay$mes,month.abb)

ggplot(delay, aes(dist, delay)) +
  geom_point(aes(size = count), alpha = 1/2) +
  xlim(c(0,1000)) +
  ylim(c(0,250)) +
  labs(title="Relaci?n entre distancia de los vuelos de JFK y retrasos en su salida", x="Distancia (mil
  geom_smooth(method = 'gam') +
  scale_size_area() +
  scale_radius(name="N?m. vuelos") +
  facet_wrap(~mes, ncol = 3)
```

```
## Scale for 'size' is already present. Adding another scale for 'size',
## which will replace the existing scale.
```

```
## Warning: Removed 32 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 32 rows containing missing values (geom_point).
```



Al igual que en los anteriores parece no haber una relacion muy fuerte, aunque si se aprecia algo mas una peque?a correlacion negativa. A priori con estos resultados parece que las dos variables no tengan mucha relacion aunque habria que hacer un estudio mas a fondo para poder confirmarlo.

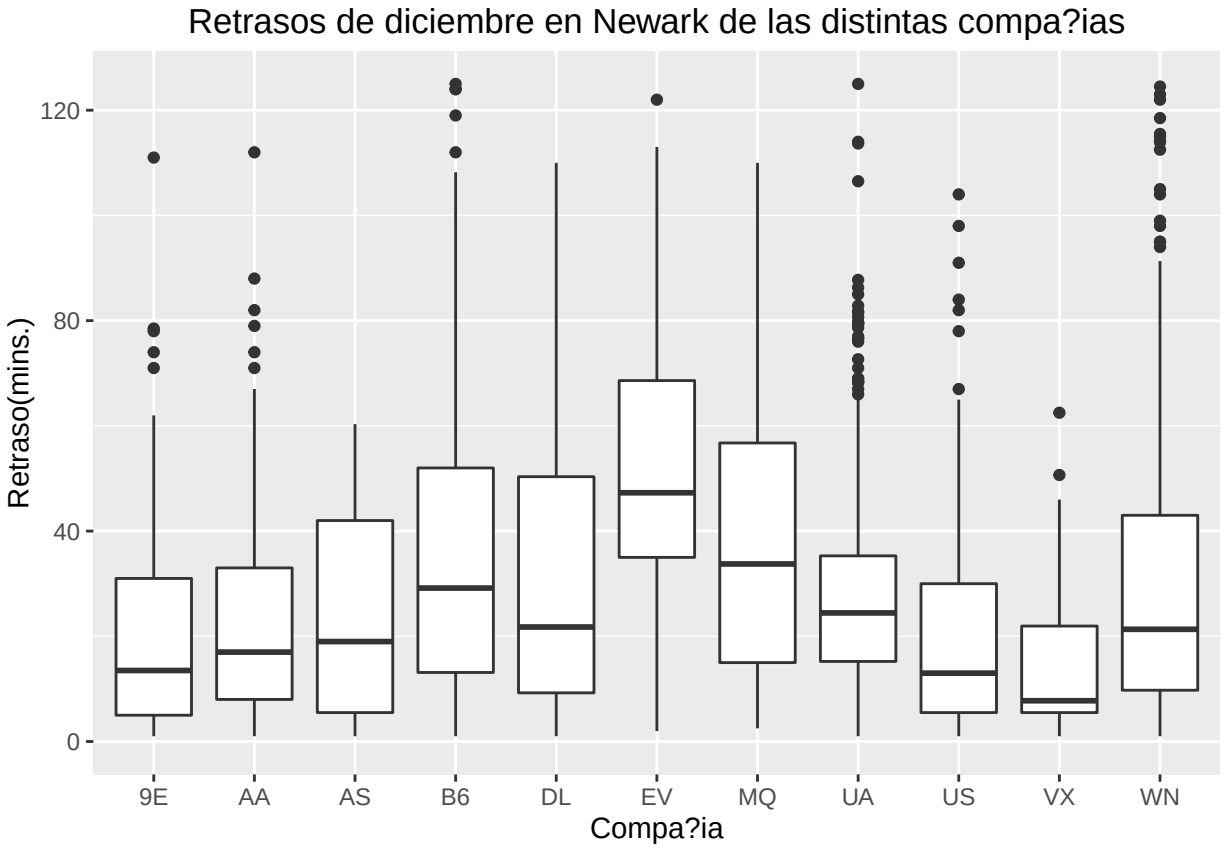
Ejercicio 4 (*20 puntos*)

Utilizando boxplots (`geom_boxplot`), representar gráficamente una comparativa de los retrasos de salida entre las distintas compañías aéreas, en el mes de diciembre, para el aeropuerto de Newark (código 'EWR'). ¿Se observan diferencias notables?

```
by_tailnum <- flights %>%
  filter(dep_delay>0, month==12, origin == 'EWR') %>%
  group_by(tailnum, carrier)
delay <- summarise(by_tailnum,
                   delay = mean(dep_delay, na.rm = TRUE))

ggplot(delay, aes(carrier, delay)) +
  geom_boxplot() +
  scale_y_continuous(limits=c(0,125)) +
  labs(title = "Retrasos de diciembre en Newark de las distintas compa?ias", x = "Compa?ia", y = "Retras
```

```
## Warning: Removed 40 rows containing non-finite values (stat_boxplot).
```



En este caso si que se pueden apreciar diferencias notables entre las distintas compa?ias. Se puede ver como la mediana varia desde cerca de los 10 minutos para el caso de VX hasta los 50 minutos para EV, que como vemos en el grafico coincide con que era la que mayores retrasos acumulaba en las salidas de sus vuelos. En cuanto a los recorridos intercuartilicos tambien se ve que son distinos teniendo algunos como MQ, DL, B6. . . los datos mas dispersos en comparaci?n con otras como VX o UA.