Máster en Data Science. URJC

Técnicas y Métodos de Ciencia de Datos

Práctica Final R

Felipe Ortega, Isaac Martín

2017

Índice

0.1 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`.

0.2 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)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
library(dplyr)
library(nycflights13)
```

0.3 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?

```
# Introduce aquí el código para generar la respuesta
# Puedes añadir más chunks si es necesario
```

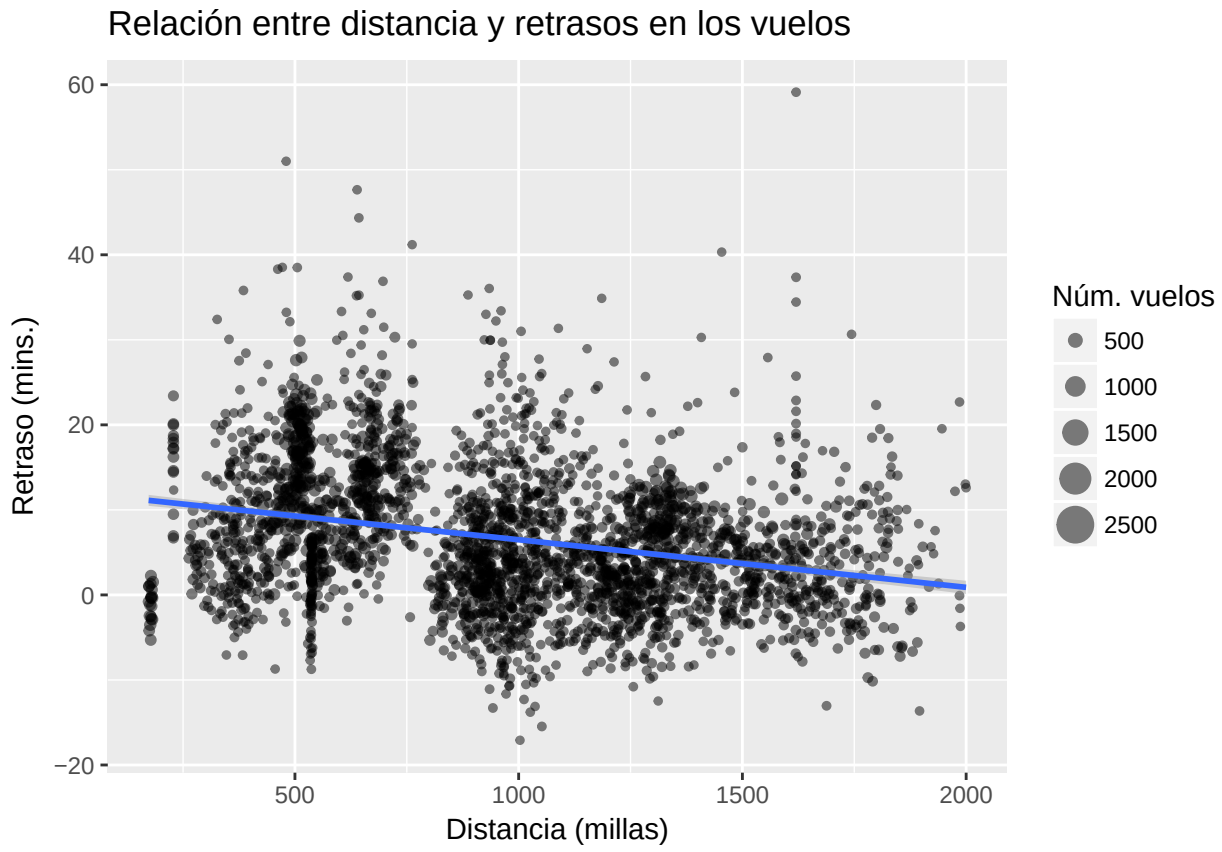
0.4 Ejercicio 2 (*30 puntos*)

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.

```
# Introduzca aquí el código para generar la respuesta
```

[INTRODUCE AQUÍ LAS CONCLUSIONES]

0.5 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?

```
# Introduzca aquí el código para generar la respuesta
```

[INTRODUCE AQUÍ LAS CONCLUSIONES]

```
# Introduzca aquí el código para generar la respuesta
```

[INTRODUCE AQUÍ LAS CONCLUSIONES]

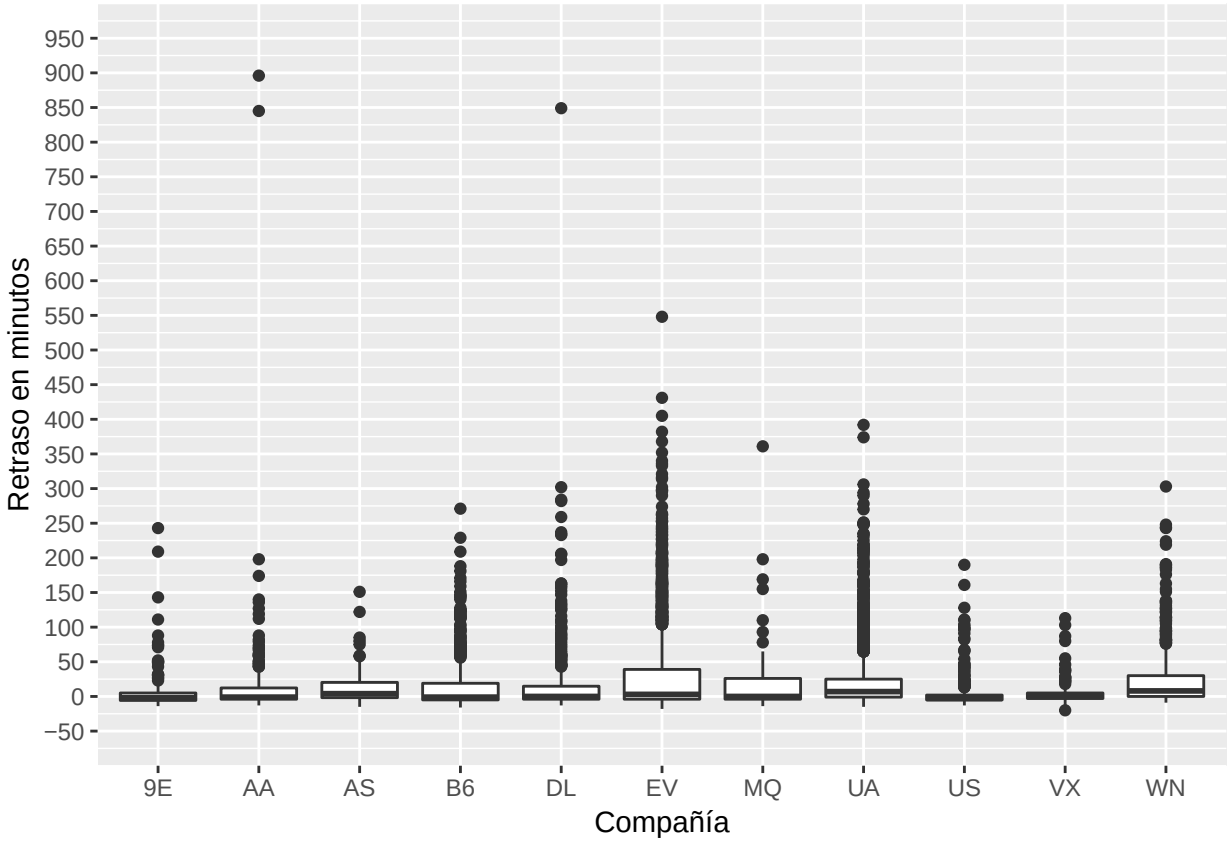
0.6 Ejercicio 4 (*20 puntos*)

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?

```
#filtramos los datos por mes, aeropuerto origen y descartamos los valores NAN del camp
flights_filtrado <- filter(flights, month==12 &
                             origin=='EWR' &
                             !is.na((dep_delay)))
#agrupamos los datos por compañía para poder mostrara acontinuación el gráfico
```

```
flights_by_carrier <-  flights_filtrado %>% group_by(carrier)

#presentamos el gráfico, fiajndo un escala que permita ver los límites del boxplot más
ggplot(flights_by_carrier, aes(x=carrier, y=dep_delay)) +
  geom_boxplot() +
  scale_y_continuous(name = "Retraso en minutos",
                     breaks = seq(-50, 950, 50),
                     limits=c(-50, 950)) +
  scale_x_discrete(name="Compañía")
```



Si bien es cierto que la mayor parte de los retrasos en todas las compañías es inferior a los 50 minutos, se observa también gran cantidad de datos atípicos con un amplio rango de valores.

La mediana, muy similar en todas ellas, indica que lo normal es que el retraso sea inferior a los 25 minutos.