Primer Parcial - Modelo B

Análisis de Datos

2022-10-27

Introducción

El paquete `nycflights13` contiene información sobre todos los vuelos que partieron de la ciudad de Nueva York (NYC) de los aeropuertos Aeropuerto Internacional Libertad de Newark (EWR), Aeropuerto Internacional John F. Kennedy (JFK) y Aeropuerto Internacional de La Guardia LGA en 2013. Los datos se encuentran en el dataframe `flights`.

Ejecuta el siguiente bloque de código para cargar las librerías necesarias para el examen, y crear los datos `flights_small`, que son una versión reducida del conjunto de datos `flights` con únicamente 10000 vuelos. Para el resto del examen, utiliza los datos `flights_small`.

```
library(tidyverse)
library(nycflights13)

set.seed(13)
flights_small <- flights %>% sample_n(10000)
```

Alternativamente, si lo anterior no te funciona, puedes descargar los datos `flights_small.csv` de Canvas y colocarlos en la carpeta donde estés ejecutando R. Una vez hayas hecho esto, puedes leerlos usando

```
flights_small <- read_csv("flights_small.csv")
```

Esta base de datos contiene información acerca de 19 variables. No obstante, para este examen únicamente utilizaremos las siguientes:

- `hour`: hora del vuelo.
- `month`: mes del vuelo.
- `origin`: aeropuerto de origen.
- `dest`: aeropuerto de destino.
- `air_time` : tiempo en el aire, en minutos.
- `carrier`: aerolínea
- `dep_delay`: retraso en despegue (en minutos).
- `arr_delay`: retraso en la llegada (en minutos).

Ejercicio 1 (2 pts)

Para los vuelos que partieron de los aeropuertos de Nueva York en 2013, ¿cuáles son las siete combinaciones de aeropuertos de origen y destino más comunes? Presenta los resultados en términos de la proporción sobre el total de vuelos que aparecen en los datos. Para las siete combinaciones más comunes, muestra **únicamente** el aeropuerto de origen, el aeropuerto de destino, la hora del vuelo, y la proporción correspondiente en su tabla.

Aclaración: Si entre los aeropuertos X e Y hubiese 3000 vuelos en los datos, la proporción que se pide es 0.3, ya que los datos tienen 10000 registros en total.

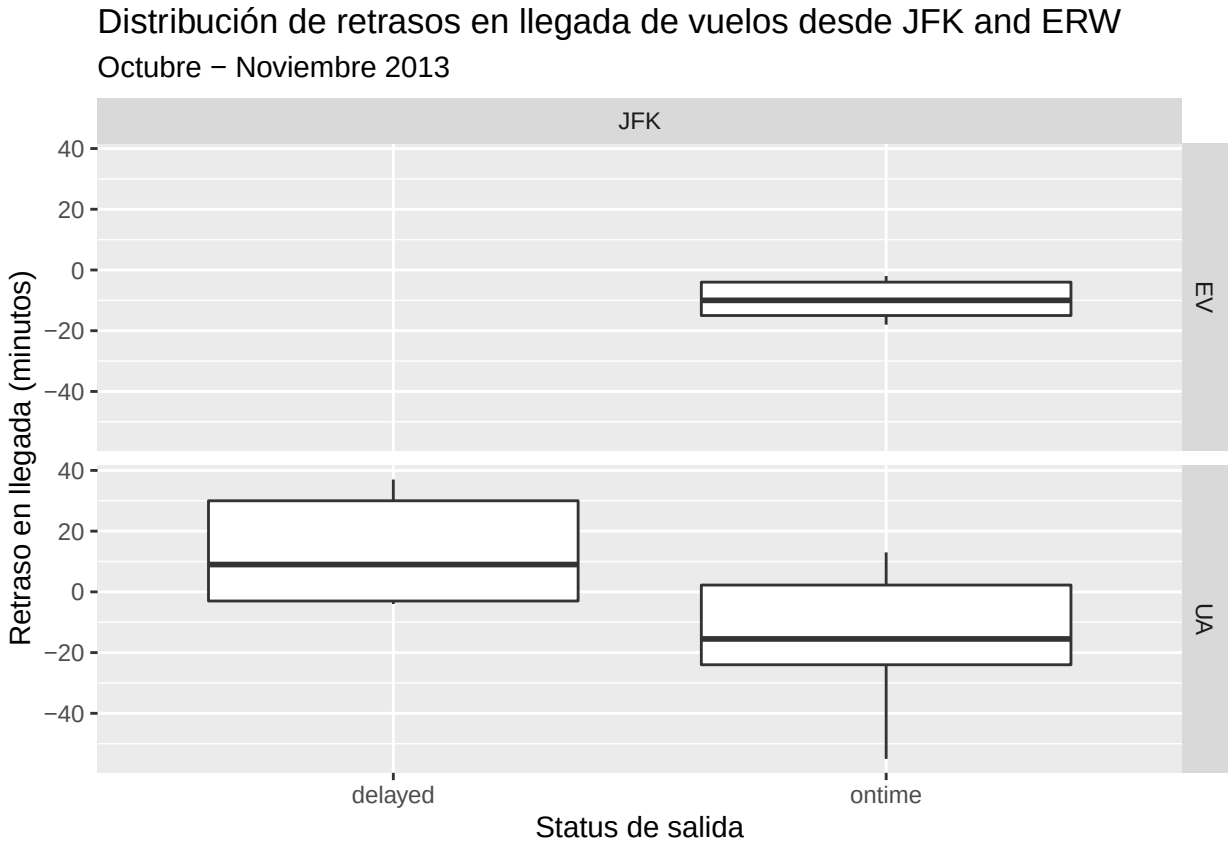
Ejercicio 2 (1 pts)

¿Cuáles fueron las cinco horas del día en las que existen más vuelos? Haz una tabla en la que aparezca **únicamente** las **cinco** horas con más vuelos en orden **descendente** de número de vuelos y muestre también la cantidad de vuelos que despegaron durante esas horas. (Ten en cuenta que las horas están escritas en formato de 24 horas, por ejemplo, `hour` = 14 son las 2 p. m.).

Ejercicio 3 (5 pts)

Recrea el siguiente gráfico utilizando los datos de `flights_small`. Fíjate en todos los detalles. Además, describe el propósito de esta visualización y extrae alguna conclusión.

- Pista 1: la visualización usa una variable llamada `dep_status`, que no está incluida en los datos `flights_small`. Tendrá que crear `dep_status` tú mismo mismo. Se trata de una variable categórica que es igual a "ontime" cuando `dep_delay` <= 0 y "delayed" cuando `dep_delay` > 0.*
- Pista 2: EV y UA son dos aerolíneas. JFK y ERW son los aeropuertos de origen.
- Pista 3: elimina todas las filas en las que `arr_delay` **y** `dep_delay` tengan un valor ausente.



Ejercicio 4 (2 pts)

Considere solo los vuelos que **no** tienen un na en la información de tiempo en el aire: ¿Los vuelos a qué aeropuerto tienen la media de tiempo en el aire más alta? ¿Cuánto vale esta media? ¿Es esperable este resultado?

*Pista: si no conoces las siglas del aeropuerto resultante, puedes usar internet para determinarlas.