Examen 1: Filtrado y Manipulación de Datos

Jesus Mudarra Luján

2022-11-13

Pregunta 1

Intenta describir con frases entendibles el conjunto de vuelos retrasados. Intenta dar afirmaciones como por ejemplo:

- Un vuelo tiende a salir unos 20 minutos antes el 50% de las veces y a salir tarde el 50% de las veces restantes.
- Los vuelos de la compañía XX llegan siempre 20 minutos tarde.
- El 95% de los vuelos a HNL llegan a tiempo, pero el 5% restante se retrasan más de 3 horas.

Pregunta 2

Da una versión equivalente a las pipes siguientes sin usar la función `count`:

```
not_cancelled %>% count(dest)
not_cancelled %>% count(tailnum, wt = distance)
```

Pregunta 3

Para definir un vuelo cancelado hemos usado la función:

```
(is.na(dep_delay) | is.na(arr_delay))
```

Intenta dar una definición que sea mejor, ya que la nuestra es un poco subóptima. ¿Cuál es la columna más importante?

Pregunta 4

Investiga si existe algún patrón del número de vuelos que se cancelan cada día.

Investiga si la proporción de vuelos cancelados está relacionada con el retraso promedio por día en los vuelos.

Investiga si la proporción de vuelos cancelados está relacionada con el retraso promedio por aeropuerto en los vuelos.

¿Qué compañía aérea sufre los peores retrasos?

Pregunta 5

Difícil: Intenta desentrañar los efectos que producen los retrasos por culpa de malos aeropuertos vs malas compañías aéreas. Por ejemplo, intenta usar:

```
flights %>% group_by(carrier, dest) %>% summarise(n())
```

Pregunta 6

¿Qué hace el parámetro `sort` como argumento de `count()`? ¿Cuándo puede sernos útil?

Vuelve a la lista de funciones útiles para filtrar y mutar y describe cómo cada operación cambia cuando la juntamos con un `group_by`.

Pregunta 7

Vamos a por los peores aviones. Investiga el top 10 de qué aviones (número de cola y compañía) llegaron más tarde a su destino.

Pregunta 8

Queremos saber qué hora del día nos conviene volar si queremos evitar los retrasos en la salida.

Difícil: Queremos saber qué día de la semana nos conviene volar si queremos evitar los retrasos en la salida.

Pregunta 9

Para cada destino, calcula el total de minutos de retraso acumulado.

Para cada uno de ellos, calcula la proporción del total de retraso para dicho destino.

Pregunta 10

Los retrasos suelen estar correlacionados con el tiempo. Aunque el problema que ha causado el primer retraso de un avión se resuelva, el resto de vuelos se retrasan para que salgan primero los aviones que debían haber partido antes. Intenta usar la función `lag()` y explora cómo el retraso de un avión se relaciona con el retraso del avión inmediatamente anterior o posterior.

Pregunta 11

Vamos a por los destinos esta vez. Localiza vuelos que llegaron 'demasiado rápido' a sus destinos. Seguramente, el becario se equivocó al introducir el tiempo de vuelo y se trate de un error en los datos. Calcula para ello el cociente entre el tiempo en el aire de cada vuelo relativo al tiempo de vuelo del avión que tardó menos en llegar a dicho destino. ¿Qué vuelos fueron los que más se retrasaron en el aire?

Pregunta 12

Encuentra todos los destinos a los que vuelan dos o más compañías y para cada uno de ellos, crea un ranking de las mejores compañías para volar a cada destino (utiliza el criterio que consideres más conveniente como probabilidad de retraso, velocidad o tiempo de vuelo, número de vuelos al año...)

Finalmente, para cada avión (basándonos en el número de cola) cuenta el número de vuelos que hace antes de sufrir su primer retraso de más de una hora. Valora entonces la fiabilidad del avión o de la compañía aérea asociada al mismo.