Deber 1
Estadística

Federico Zertuche

Importante: Escriban el código R para TODAS las preguntas de programación. Cuando terminen suban un pdf con las respuestas a la sección correspondiente del aula virtual.

Ejercicio 1 Un poco de R.

Pregunta 1 *Cuáles de las siguientes expresiones valen 99 para x = 10 en R? Analicen la sintaxis como si estuvieran programando.*

- *10x - 1*
- *(x)(x) - 1*
- *abs(x*x) - abs(9-x)*
- *11 * x - x + 1*

Pregunta 2 *Un vector contiene una serie de ganancias ordenadas de manera creciente. Escriban el código que genera:*

- *La suma de todas las ganancias.*
- *La segunda ganancia mas grande.*
- *La diferencia mas grande entre las ganancias.*

- *Un booleano que responda a la pregunta: La mas grande diferencia ente dos ganancias es mayor a 10?*
- *La menor diferencia positiva entre dos ganancias.*
- *El máximo número de ganancias que pueden sumar sin pasar de 10000.*

Ejercicio 2 Dplyr en los Aeorpuertos.

Vamos a estudiar los datos de los vuelos locales en Estados Unidos durante el 2011. Usen los verbos:

- select()
- filter()
- mutate()
- arrange()
- group_by()
- summarise()

para manipular 3 data frames: Uno con los vuelos (flights), uno con los aviones (planes) y otro con el clima (weather). La descripción de todos los data frames está en la documentación del paquete nycflights13 (https://cran.r-project.org/web/packages/nycflights13/nycflights13.pdf).

Pregunta 3 *Instalen la librería nycflights13. Escriban el código para encontrar todos los vuelos que:*

- *Fueron de JKF(John F. Kennedy) hasta OAK(Oakland).*
- *Salieron en Enero.*

- *Tienen demoras de mas de una hora (las demoras están en minutos).*
- *Salieron entre medianoche y las 5 a.m.*
- *Tuvieron una demora de llegada 2 veces mas grande que la de salida.*

Pregunta 4 *Lean la ayuda de select(). Escriban 2 formas de selecionar las dos variables de retraso.*

Pregunta 5 *Ordenen la tabla por fecha de salida y tiempo. Cuáles fueron los vuelos que sufrieron las mayores demoras? Cuáles recuperaron la mayor cantidad de tiempo durante el vuelo?*

Pregunta 6 *Calculen la velocidad en mph usando el tiempo (que está en minutos) y la distancia (que está en millas). Cuál fué el avión que voló mas rápido?*

Pregunta 7 *En dplyr el comando pipeline %>% se lee entonces. Significa:*

$$x \,\%>\%\, f(y) \longrightarrow f(x, y).$$

Es decir pasa x como primer argumento de f.

Qué significan las siguientes líneas de código:

$$\begin{aligned} flights \,\%>\%\, & filter(!\, is.na(dep\_delay)) \,\%>\%\\ & group\_by(date,\ hour) \,\%>\%\\ & summarise(delay = mean(dep\_delay),\ n = n()) \,\%>\%\\ & filter(n > 10) \end{aligned}$$

Pregunta 8 *Cuál es la destinación que tiene las demoras promedio mas grandes? Cuántos vuelos diarios hay? Cuál es la mejor hora para viajar sin retraso?*

Ejercicio 3 6,000 Años de Urbanización Global.

En el artículo Spatializing 6,000 years of global urbanization from 3700 BC to AD 2000 los autores reunen una serie de datos sobre la población de las ciudades del mundo desde 3700 A.C. hasta el 2000 D.C. Usemos estos datos para tratar de entender algunas cosas sobre la distribución espacial de las personas a través del tiempo.

Para hacer esto vamos a tener que manipular dos data frames - uno para la población y otro con los nombres de los países y continentes - usando dplyr para eventualmente unirlos.

En este ejercicio vamos a usar algunos verbos para unir data frames.

- inner_join(d1, d2) contiene solo las filas de d1 comunes a d2.
- left_join(d1, d2) contiene todas las filas de d1 y NA en las filas de d2 que no están en d1.
- semi_join(d1, d2) no añade columnas a d1. Contiene las filas de d1 comunes a d2.

Pregunta 9 *Cuántas personas han habitado Sur América? Hay alguna forma de usar los datos para saber cuál es la región que tiene las poblaciones mas antiguas? Pueden usar los datos para trata r de entender si hay un patrón migratorio a lo largo de la historia?*