dw-2020-parcial-1

Marianna Flores
9/9/2020

Examen parcial

Indicaciones generales:

- Usted tiene el período de la clase para resolver el examen parcial.
- La entrega del parcial, al igual que las tareas, es por medio de su cuenta de github, pegando el link en el portal de MiU.
- Pueden hacer uso del material del curso e internet (stackoverflow, etc.). Sin embargo, si encontramos algún indicio de copia, se anulará el exámen para los estudiantes involucrados. Por lo tanto, aconsejamos no compartir las agregaciones que generen.

Sección I: Preguntas teóricas.

- Existen 10 preguntas directas en este Rmarkdown, de las cuales usted deberá responder 5. Las 5 a responder estarán determinadas por un muestreo aleatorio basado en su número de carné.
- Ingrese su número de carné en `set.seed()` y corra el chunk de R para determinar cuáles preguntas debe responder.

```
set.seed(20180040)
v<- 1:10
preguntas <-sort(sample(v, size = 5, replace = FALSE ))

paste0("Mis preguntas a resolver son: ",paste0(preguntas,collapse = ", "))
```

```
## [1] "Mis preguntas a resolver son: 2, 3, 4, 5, 9"
```

Preguntas Teoricas

1. ¿Por qué en R utilizamos funciones de la familia apply (lapply,vapply) en lugar de utilizar ciclos? Las funciones de R en de las librerias de lapply y vapply son mas comunes para el analisis de datos por lo que familiarizarse con el uso es altamente conveniente. Por otra parte la programacion de los ciclos puede ser compleja y que el codigo se vea menos limpio es decir que sea mas complicado manejarlo.
2. ¿Cuál es la forma correcta de cargar un archivo de texto donde el delimitador es ：? Para cargar un archivo con un delimitador especifico se debe establecer cual es destro de la funcion por ejemplo df <- read_delim("~/nombre_archivo.txt", delim = ":") Siempre es conveniente saber cual es el delimiter default de

la computadora e incluso de la region para despues no tener inconvenientes y saber cuando se debe especificar como en el caso de ":"

3. ¿Qué pasa si quiero agregar una nueva categoría a un factor que no se encuentra en los niveles existentes? Se debe de agregar al final de los niveles ya existentes o especificar que nivel se va a modificar por esta nueva variable. Dependiendo del tipo de estructura al que se esta agregando esta nueva categoria/unidad, si el sistema lo puede procesar. En el caso un vector no se puede agregar una nueva categoria de un nivel que no se tiene, en otras estructuras si es posible como en un string.

4. Si en un dataframe, a una variable de tipo `factor` le agrego un nuevo elemento que *no se encuentra en los niveles existentes*, ¿cuál sería el resultado esperado y por qué?
   - El nuevo elemento
   - `NA` Si se agrega a una nueva celda/variable un elemento de un tipo distinto al que ya habia en el data frame (digamos texto a una columna de factor) se va a guardar el nuevo elemento nuevo. Por ser un dataframe cada variable funciona por aparte, en el caso de un vector si daria un error, pero por ser df se agrega con normalidad a los datos que ya se tienen.

5. Si quiero obtener como resultado las filas de la tabla A que no se encuentran en la tabla B, ¿cómo debería de completar la siguiente sentencia de SQL?

   - SELECT * FROM A LEFTJOIN B ON A.KEY = B.KEY WHERE B.KEY IS NULL

Extra: ¿Cuántos posibles exámenes de 5 preguntas se pueden realizar utilizando como banco las diez acá presentadas? Hay 252 posibilidades

```
cantidad <- combn(v,5, simplify = FALSE)
```

Sección II Preguntas prácticas.

- Conteste las siguientes preguntas utilizando sus conocimientos de R. Adjunte el código que utilizó para llegar a sus conclusiones en un chunk del markdown.

A

De los clientes que están en más de un país,¿cuál cree que es el más rentable y por qué?

```
library(dplyr)
library(readr)
library(tidyverse)
library(readxl)

df <-read_rds("~/Data Wrangling/Ejercicios/Parcial1/parcial_anonimo.rds")

clientes_totales <- df %>%
  select(Pais, Cliente, Venta) %>%
  group_by(Cliente, Pais) %>%
  summarise(ventas = sum(Venta), .groups = 'drop') %>%
  arrange(desc(ventas))

clientes_totales
```

```
## # A tibble: 2,154 x 3
##    Cliente  Pais      ventas
##    <chr>    <chr>      <dbl>
##  1 af267306 4f03bd9b 393666.
##  2 f6e6ba91 4f03bd9b 185830.
##  3 93730e73 4f03bd9b 135499.
##  4 9314226b 4f03bd9b 126636
##  5 f217abbd 4f03bd9b 112183.
##  6 e1123460 4f03bd9b  80362.
##  7 dd0d7f4d 4f03bd9b  80288.
##  8 f8e43f3e 4f03bd9b  79406.
##  9 09c918d8 4f03bd9b  74837.
## 10 e1d1f5f7 4f03bd9b  62455.
## # ... with 2,144 more rows
```

El cliente mas rentable es f217abbd *Lo trate de calcular manual porque ya no tenia tiempo, por eso no hay mas proceso

B

Estrategia de negocio ha decidido que ya no operará en aquellos territorios cuyas pérdidas sean "considerables". Bajo su criterio, ¿cuáles son estos territorios y por qué ya no debemos operar ahí?

```
total <- df %>% summarise(sum(Venta))

territorios <- df %>%
  select(Territorio, Venta, 10) %>%
  group_by(Territorio) %>%
  summarise(ingresos = sum(Venta),porcentaje = ingresos/6286229, .groups = 'drop') %>% arrange(ingresos)

territorios
```

```
## # A tibble: 104 x 3
##    Territorio ingresos porcentaje
##    <chr>         <dbl>      <dbl>
##  1 e6fd9da9       18.2 0.00000289
##  2 13b223c9       49.9 0.00000794
##  3 368301e2      121.  0.0000193
##  4 79428560      132   0.0000210
##  5 e034e3c8      247.  0.0000393
##  6 0bfe69a0      384.  0.0000611
##  7 456278b8      493.  0.0000783
##  8 4163fa3f      580.  0.0000922
##  9 3e0d75d0      647.  0.000103
## 10 aed8e579      747.  0.000119
## # ... with 94 more rows
```

En este momento se desconocen los costos totales que indican para la empresa las ventas, La mayoria (casi 50%) de los territorios vende mas de 10000 por lo que se considera prudente dejar esos territorios que no representan ni un 1%