dw-2020-parcial-1

Emily Soto

20/09/2021

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.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1
```

```r
set.seed(20190508)
v<- 1:10
preguntas <-sort(sample(v, size = 6, replace = FALSE ))

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

```
## [1] "Mis preguntas a resolver son: 1, 2, 3, 5, 7, 9"
```

Listado de preguntas teóricas

1. Para las siguientes sentencias de `base R`, liste su contraparte de `dplyr`:

* ‘str()‘ R// glimpse()
* ‘df[,c("a","b")]‘ R// df %>% select(a,b)
* ‘names(df)[4] <- "new_name"‘ donde la posición 4 corresponde a la variable ‘old_name‘
  R// df %>% rename(new_name=‘old name‘)
* ‘df[df$variable == "valor",]‘
  R// df %>% filter(variable=="valor")

2. Al momento de filtrar en SQL, ¿cuál keyword cumple las mismas funciones que el keyword `OR` para filtrar uno o más elementos una misma columna?

R// El keyword "IN"

3. ¿Por qué en R utilizamos funciones de la familia apply (lapply,vapply) en lugar de utilizar ciclos?

- Es más eficiente, ya que cruza los datos de multiples maneras en lugar de asignar un valor a cada va
- Se entiende mejor el código: a primera vista, no hay que tratar de entender como funciona un ciclo
-El código es más adaptable a otras variables o dataset

5. ¿Cuál es la forma correcta de cargar un archivo de texto donde el delimitador es :?

R// con la librería "readr" -> read_delim("df", delim = ":")

7. ¿Qué pasa si quiero agregar una nueva categoría a un factor que no se encuentra en los niveles existentes?

Si se desea agregar la categoría y no existe ya dentro de los niveles, R asignará un NA dentro de los

```r
genero <- c("h", "m")
x=c("Otro","h","m")
factor(x, levels = genero)
```

```
## [1] <NA> h    m
## Levels: h m
```

Sin embargo, si se agrega como un nuevo nivel si lo hará sin problemas con la función levels() o releve

9. En SQL, ¿para qué utilizamos el keyword `HAVING`?

"HAVING" es una función que funciona como un WHERE, pero a diferencia de esta, se pueden hacer agrega

```
library(gtools)
```

Extra: ¿Cuántos posibles exámenes de 5 preguntas se pueden realizar utilizando como banco las diez acá presentadas?(responder con código de R.)

```
## Warning: package 'gtools' was built under R version 4.1.1
```

```
preguntas <- c(1:10)
combinaciones <- combinations(10, 5, preguntas)
nrow(combinaciones)
```

```
## [1] 252
```

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

*Se asume que la variable "venta" ya está en utilidad total por pedido De los clientes que están en más de un país,¿cuál cree que es el más rentable y por qué?

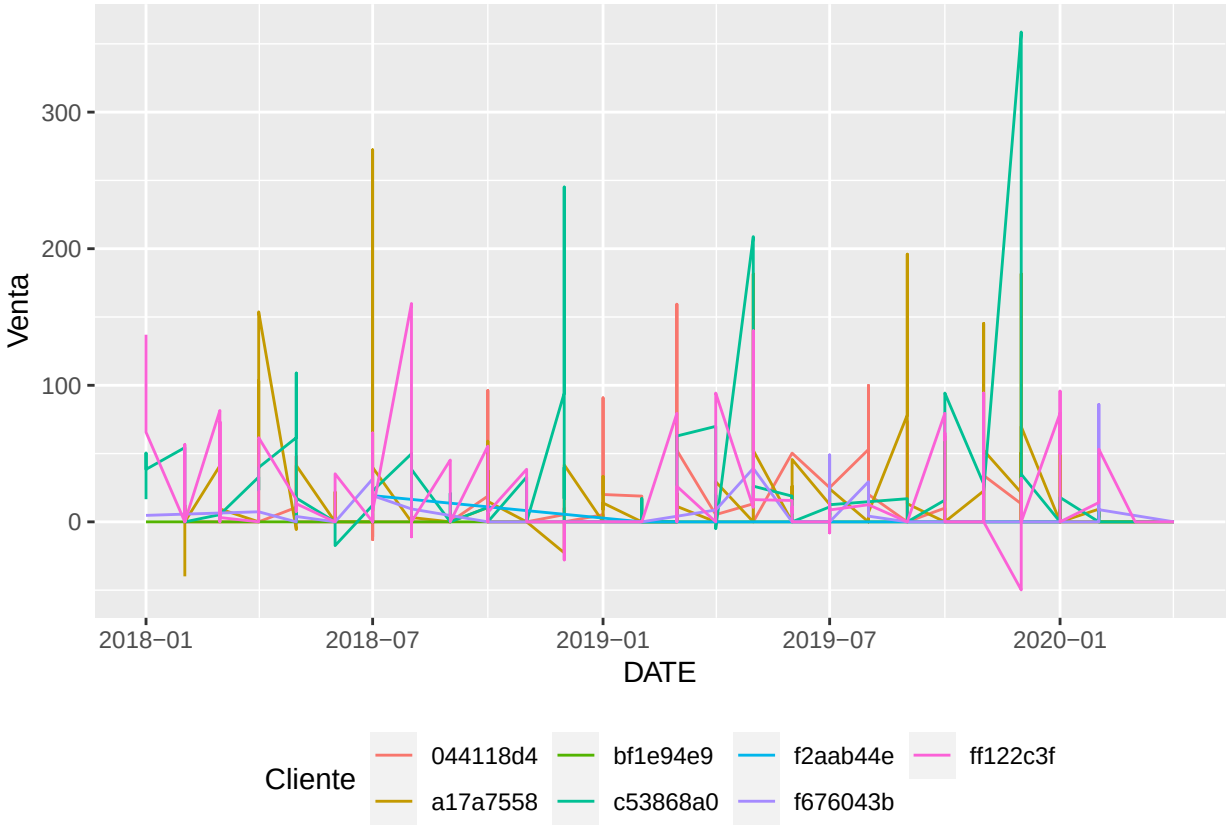
```
#Resolución
df=readRDS("C:/Users/Emily Soto/Desktop/Data Science/DW/Parcial1/parcial_anonimo.rds")
clientes_paises=df %>%select(Cliente, Pais, `Unidades plaza`, Venta) %>% group_by(Cliente) %>% summaris
clientes_paises[order(clientes_paises$Venta, decreasing = T),]
```

```
## # A tibble: 7 x 5
##   Cliente  Pais_presencia Unidades  Venta Proporcion
##   <chr>             <int>    <dbl>  <dbl>      <dbl>
## 1 a17a7558              2     2274 19818.     0.115
## 2 ff122c3f              2     1363 15359.     0.0887
## 3 c53868a0              2     1690 13813.     0.122
## 4 044118d4              2     1134  9436.     0.120
## 5 f676043b              2      377  3635.     0.104
## 6 f2aab44e              2       35   400.     0.0874
## 7 bf1e94e9              2        0     0    NaN
```

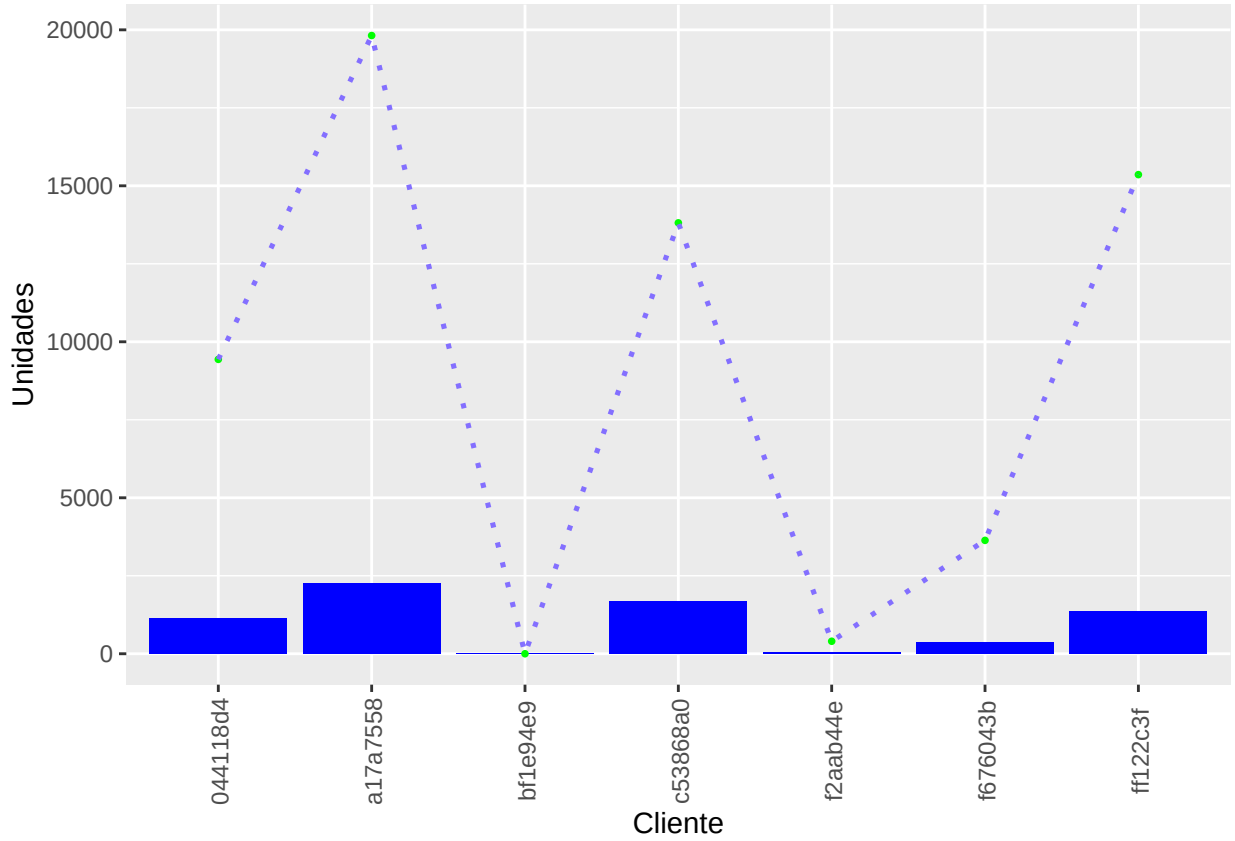
```
clientes=clientes_paises$Cliente
```

```
ggplot(data=df[df$Cliente==clientes,], aes(x=DATE, y=Venta, group=Cliente))+ geom_line(aes(color=Client
```

```
## Warning in df$Cliente == clientes: longitud de objeto mayor no es múltiplo de la
## longitud de uno menor
```




```
ggplot(data=clientes_paises)+
  geom_bar(aes(x=Cliente,y=Unidades), fill='blue', stat="identity") +
    geom_point(aes(x=Cliente,y=Venta), color = rgb(0, 1, 0), pch=16, size=1) +
    geom_path(aes(x=Cliente,y=Venta, group=1), colour="slateblue1", lty=3, size=0.9)+theme(axis.text.x =
```



Dentro de los clientes más rentables, basados en el criterio de ventas y unidades, se encuentran los clientes: a17a7558, ff122c3f y c53868a0. Además, en las gráficas se puede notar que no solo tienen un buen performance a través del tiempo sino que tambien podemos ver que la utilidad es bastante alta en porporción a las pocas unidades vendidas.

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

```
###resuelva acá
summary(df$Venta)
```

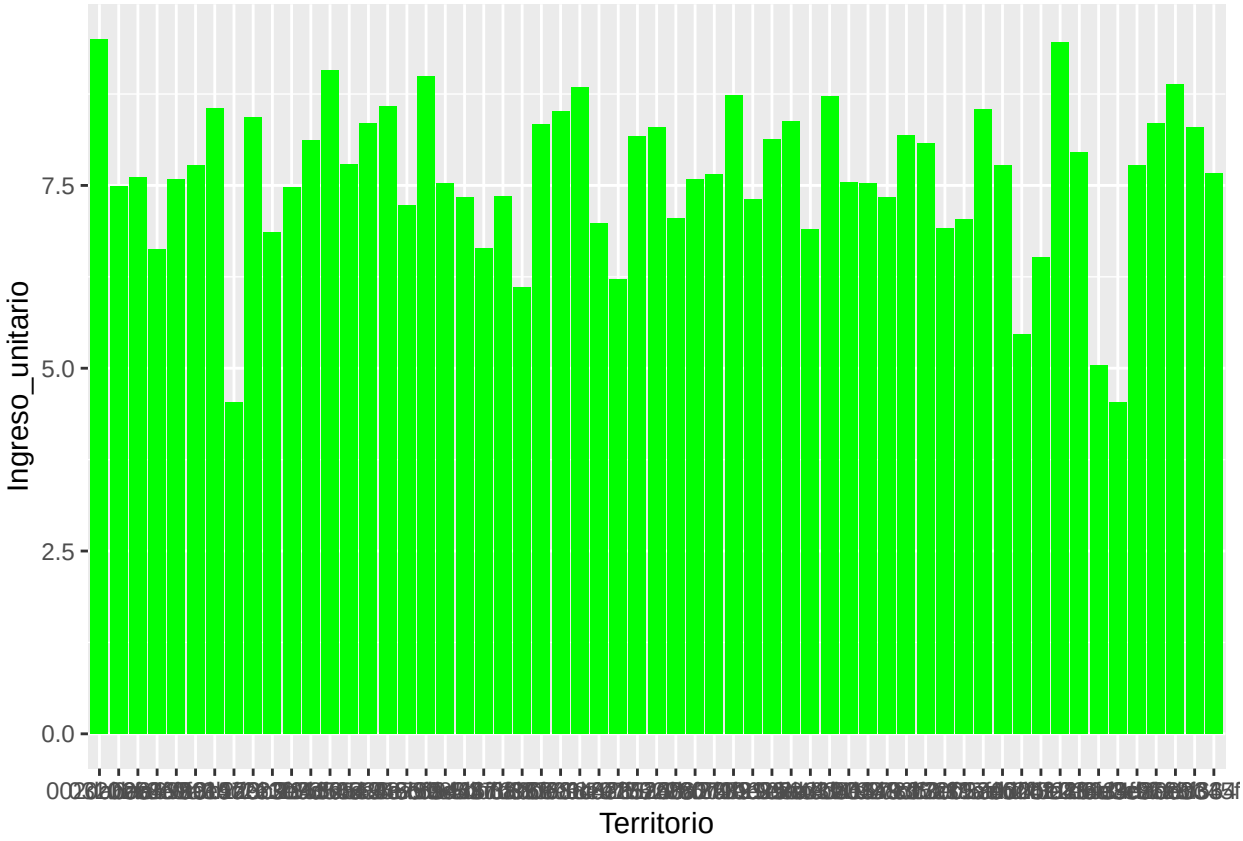
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##  -622.60      0.00      0.00     27.79     23.31 44460.00
```

```
bajo_media=df %>%select(Territorio,  `Unidades plaza`, Venta) %>% group_by(Territorio) %>% summarise(Un
bajo_media=bajo_media[order(bajo_media$Ingreso_unitario, decreasing = F),]
bajo_media
```

```
## # A tibble: 59 x 4
##    Territorio Unidades  Ventas Ingreso_unitario
##    <chr>         <dbl>   <dbl>            <dbl>
##  1 13b223c9         11    49.9             4.54
##  2 e6fd9da9          4    18.2             4.54
```

```
##  3 e034e3c8         49   247.              5.04
##  4 cf970512       1166  6375.              5.47
##  5 67696f68       7734 47176.              6.10
##  6 6c8335a4        578  3594.              6.22
##  7 d02bf225        202  1315.              6.51
##  8 0bfe69a0         58   384.              6.63
##  9 5d43dd39        976  6479.              6.64
## 10 1a9b2b4c        939  6437.              6.86
## # ... with 49 more rows
```

```
ggplot(data=bajo_media)+
  geom_bar(aes(x=Territorio, y=Ingreso_unitario), fill='green', stat="identity")
```



Los territorios que se deben cerrar, basados en el criterio de ingreso marginal, son aquellos territorios con un ingreso menor a 7 por unidad, considerando que la media está por encima de 9.5. Estos son:

```
bajo_media[1:13,1]
```

```
## # A tibble: 13 x 1
##    Territorio
##    <chr>
##  1 13b223c9
##  2 e6fd9da9
##  3 e034e3c8
##  4 cf970512
##  5 67696f68
```

```
##  6 6c8335a4
##  7 d02bf225
##  8 0bfe69a0
##  9 5d43dd39
## 10 1a9b2b4c
## 11 9de43341
## 12 c072f75a
## 13 68de9759
```