Reporte-Segmentación RFM

20/01/2022

Contents

INSTRUCCIONES **2**

PASO 1 **2**

PASO 2 **3**

PASO 3 **4**

PASO 4 **6**

PASO 5 **6**

INSTRUCCIONES

En esta actividad tu habilidad para realizar un análisis RFM será evaluado, así como tu criterio para responder las preguntas y presentar un análisis formal.Ante todo deben cargar las librerias que necesitarán para trabajar: readr, dplyr y caret

```
# La función {library} sirve para cargar los paquetes,
# los cuales debieron ser preiamente instalados con la función {install packages}.
# La instalación de paquetes se realiza solo la primera vez,
# pero se deben cargar los paquetes cada vez que se inicia el programa

library(readr)
library(dplyr)
library(caret)
```

PASO 1

Importar el archivo csv llamado "dataset_Actividad1.csv" y realizar un análisis exploratorio de la información utilizando al menos 3 funciones.

¿Qué puedes mencionar acerca del data set de trabajo? Expresa tus ideas en un mínimo de 5 líneas (1.5pts).

```
## Importar data
dataRFM <- read_csv("dataset_Actividad1.csv")
```

Table 1: Data de trabajo (continued below)

COMPANY_CD_FSK	CUSTOMER_CD_FSK	PRODUCT_CD_FSK	INVOICE_YEAR_MONTH
127	20751556	77269453	2021-10-01
127	4462	510666	2021-03-01
127	3584	512136	2021-01-01
127	4462	508388	2021-10-01
127	4176	508388	2021-03-01
127	4462	505145	2021-04-01

QUANTITY_PIECES	GROSS_SALES
24	14964
48	14943
1344	14933
96	14881
96	14853
40	14677

```
## Explorar data
View(dataRFM)
glimpse(dataRFM)
summary(dataRFM)
```

Table 3: Filas y columnas

9843	6

Table 4: Resumen de la data (continued below)

COMPANY_CD_FSK	CUSTOMER_CD_FSK	PRODUCT_CD_FSK
Min. : 72.0	Min. : 469	Min. : 152
1st Qu.:127.0	1st Qu.: 2656	1st Qu.: 497832
Median :127.0	Median : 4216	Median : 507212
Mean :147.7	Mean : 3039890	Mean : 12551233
3rd Qu.:200.0	3rd Qu.: 3469616	3rd Qu.: 12995479
Max. :241.0	Max. :32357561	Max. :154574141

INVOICE_YEAR_MONTH	QUANTITY_PIECES	GROSS_SALES
Min. :2021-01-01 00:00:00	Min. :-1000.0	Min. :-3027.6
1st Qu.:2021-03-01 00:00:00	1st Qu.: 6.0	1st Qu.: 90.6
Median :2021-06-01 00:00:00	Median : 20.0	Median : 315.4
Mean :2021-06-03 06:23:09	Mean : 210.5	Mean : 948.1
3rd Qu.:2021-09-01 00:00:00	3rd Qu.: 80.0	3rd Qu.: 953.5
Max. :2021-12-01 00:00:00	Max. :30000.0	Max. :14963.5

Interpretación: El archivo está compuesto por cerca de 10000 registros (filas) y 6 variables (columnas). Estas variables brindan información acerca de: los identificadores de la empresa, consumidor y producto, fecha de la facturación (compras realizadas desde enero hasta diciembre del año 2021),unidades compradas y, finalmente, sobre el precio bruto. Adicionalmente, se puede indicar que las variables de cantidades compradas y ventas brutas presentan una alta variabilidad al contar con valores máximos y mínimos muy distantes entre sí, además que contienen datos atípicos (valores en negativo).

PASO 2

Crea un nuevo data frame transformando la información original para calcular las métricas de Recency, Frenquency y Monetary Value sobre transacciones válidas y evaluando la necesidad de filtrar outliers. (1.5pts).

```
# Hay valores negativos en las variables de precio bruto y unidades, # por lo que será necesario elimin

# Eliminamos aquellos registros (filas) que no son útiles para nuestros análisis
dataRFM_clean <- dataRFM %>%
  filter(QUANTITY_PIECES > 0 & GROSS_SALES > 0)
```

```
# Verificamos que ya no existan valores atípicos
summary(dataRFM_clean)
```

Table 6: Resumen de la data limpia (continued below)

COMPANY_CD_FSK	CUSTOMER_CD_FSK	PRODUCT_CD_FSK
Min. : 72.0	Min. : 469	Min. : 152
1st Qu.:127.0	1st Qu.: 2656	1st Qu.: 497832
Median :127.0	Median : 4216	Median : 507034
Mean :147.7	Mean : 2953390	Mean : 12091967
3rd Qu.:200.0	3rd Qu.: 3469616	3rd Qu.: 12989660
Max. :241.0	Max. :32357561	Max. :154574141

INVOICE_YEAR_MONTH	QUANTITY_PIECES	GROSS_SALES
Min. :2021-01-01 00:00:00	Min. : 2.0	Min. : 0.52
1st Qu.:2021-03-01 00:00:00	1st Qu.: 6.0	1st Qu.: 104.50
Median :2021-06-01 00:00:00	Median : 20.0	Median : 333.90
Mean :2021-06-03 03:31:08	Mean : 218.4	Mean : 974.41
3rd Qu.:2021-09-01 00:00:00	3rd Qu.: 86.0	3rd Qu.: 984.94
Max. :2021-12-01 00:00:00	Max. :30000.0	Max. :14963.52

```
## Normalmente la información está registrada a nivel de producto, es decir, una línea de información p
## Lo que implica que debemos entender cuántos valores únicos contienen las columnas y tener cuidado en

length(unique(dataRFM_clean$INVOICE_YEAR_MONTH)) #12 meses
length(unique(dataRFM_clean$CUSTOMER_CD_FSK))    #107 consumidores
length(unique(dataRFM_clean$PRODUCT_CD_FSK))     #2714 productos
```

```
##Establecemos formato de fecha
dataRFM_clean$INVOICE_YEAR_MONTH <- as.Date(dataRFM_clean$INVOICE_YEAR_MONTH, "%Y-%m-%d")

# Verificamos que se cambió formato de fecha (Date)
class(dataRFM_clean$INVOICE_YEAR_MONTH)
```

[1] “Date”

PASO 3

Desarrolla un nuevo análisis exploratorio y explica los principales highlights acerca de las métricas de RFM de todos los clientes en más de 5 líneas (2pts).

```
# Aplicamos fórmulas de RFM

# Cálculo de la fecha de análisis (1 día después de la última compra)
max(dataRFM$INVOICE_YEAR_MONTH) #última compra
```

[1] “2021-12-01 UTC”

```
fecha_analisis = as.Date("2021-12-02")
```

```r
## Cálculo del RFM
## considerando Monetary como la suma de dinero gastado durante el periodo
Metricas_RFM <- dataRFM_clean %>%
  group_by(CUSTOMER_CD_FSK) %>%
  summarise(Recency=as.numeric(fecha_analisis-max(INVOICE_YEAR_MONTH)),
            Frequency=length(unique(INVOICE_YEAR_MONTH)),
            Monetary_Value=sum(GROSS_SALES))
```

```r
# Eliminamos el objeto creado puesto que vamos a considerar Monetary
# como la suma de dinero gastado durante el periodo
remove(Metricas_RFM)
```

```r
## Cálculo del RFM
## considerando Monetary como el promedio de dinero gastado durante el periodo

## Para calcular el ticket promedio,
## debemos primero agrupar por consumidor y por fecha de compra,
## sumando todas las unidades vendidas y el gross sales

dataRFM_clean <- dataRFM_clean %>%
  group_by(CUSTOMER_CD_FSK, INVOICE_YEAR_MONTH) %>%
  summarise(QUANTITY_PIECES = sum(QUANTITY_PIECES),
            GROSS_SALES     = sum(GROSS_SALES))

## para luego calcular correctamente el promedio de dinero gastado
Metricas_RFM <- dataRFM_clean %>%
  group_by(CUSTOMER_CD_FSK) %>%
  summarise(Recency=as.numeric(fecha_analisis-max(INVOICE_YEAR_MONTH)),
            Frequency=length(unique(INVOICE_YEAR_MONTH)),
            Monetary_Value= mean(GROSS_SALES))
```

```r
# Realizamos un resumen de las métricas
summary(Metricas_RFM)
```

Table 8: Resumen de las métricas

CUSTOMER_CD_FSK	Recency	Frequency	Monetary_Value
Min. : 469	Min. : 1.00	Min. : 1.000	Min. : 77.11
1st Qu.: 2361	1st Qu.: 1.00	1st Qu.: 2.000	1st Qu.: 1901.01
Median : 4462	Median : 31.00	Median : 5.000	Median : 4910.00
Mean : 5426917	Mean : 88.03	Mean : 5.832	Mean : 10867.39
3rd Qu.: 3485873	3rd Qu.:123.00	3rd Qu.: 9.000	3rd Qu.: 9806.38
Max. :32357561	Max. :335.00	Max. :12.000	Max. :113615.01

Interpretación: Los clientes evaluados presentan una media de 88 Recency, lo cual significa que en promedio no han realizado compras desde hace casi 3 meses. Además, la frecuencia de compras por cliente durante el año 2021 ha sido relativamente baja, alcanzado una media de 6 productos comprados. Finalmente, se puede indicar que el valor monetario promedio ha sido medianamente alto con un valor de 10867 euros. Adicionalmente, se puede señalar que son productos con poca rotación, pero con bastante inversión en promedio de compra. Considerando que el valor mínimo de compra es 77 euros y el máximo es 113615 aproximadamente, lo más recomendable es incentivar el valor promedio de compra de los clientes

PASO 4

Aplique el algoritmo de k-means para definir una segmentación basada en RFM. Use 5 clusters para realizar una segmentación tradicional (1.5pts).

```
#Definimos el seed para asegurarnos de poder replicar el resultado
set.seed(1234)

# Le indicamoos al algoritmo que queremos obtener 5 clusters
# utilizando solo las columnas de datos que contienen las métricas de RFM
RFM_Segmentation <- kmeans(scale(Metricas_RFM[,2:4]), 5, nstart = 1)

#Incorporamos el resultado al dataset
Metricas_RFM$Cluster <- as.factor(RFM_Segmentation$cluster)

#Analizamos la distribución de los consumidores
table(Metricas_RFM$Cluster)
```

1 2 3 4 5 6 24 39 16 22

Table 9: Distribución de los consumidores

1	2	3	4	5
6	24	39	16	22

PASO 5

Analice el resultado de la segmentación y las medias de las métricas para asignar el label correcto a cada cluster: identifica a los Champions, Loyals, Promising, At risk y Churn.

¿Por qué consideras que los Champions están incluidos en el cluster que has seleccionado como Champions? (3.5pts).

```
## Primero necesitamos entender qué tipo de consumidores se han agrupado bajo cada cluster

Metricas_RFM %>%
  select(2:5) %>%
  group_by(Cluster) %>%
  summarise_all(mean)
```

Table 10: Tipo de consumidores según cluster

Cluster	Recency	Frequency	Monetary_Value
1	21.17	9.5	72285
2	26.17	3.042	6213
3	18.79	9.923	11406
4	109.6	4.438	4840
5	280.8	1.636	2624

```r
## Asignamos un nombre relevante para los equipos de marketing (PRESTAR ATENCIÓN)
Metricas_RFM <- Metricas_RFM %>%
  mutate(Segmento = case_when(Cluster == 5 ~ "inactivo",
                              Cluster == 4 ~ "en riesgo",
                              Cluster == 2 ~ "potenciales",
                              Cluster == 1 ~ "leales",
                              Cluster == 3 ~ "champions",
                              TRUE ~ "NA"))
```

```r
## Generamos un nuevo análisis de mas medias de cada métrica de RFM
## pero sobre la columna de etiqueta del cluster (segmento)
Metricas_RFM %>%
  select(2,3,4,6) %>%
  group_by(Segmento) %>%
  summarise_all(mean)
```

Table 11: Tipo de consumidores según segmento

Segmento	Recency	Frequency	Monetary_Value
champions	18.79	9.923	11406
en riesgo	109.6	4.438	4840
inactivo	280.8	1.636	2624
leales	21.17	9.5	72285
potenciales	26.17	3.042	6213

Interpretación: Los clientes pertenecientes al cluster "3" presentan un bajo valor en **Recency** lo cual significa que realizaron recientemente una compra; además presentan un alto valor de **Frecuency** lo cual indica que realizaron varias compras durante el periodo de análisis (año 2021). Por ello, según el esquema Recency-Frecuency se ubicarían dentro de la categoría **"Champions"**. Además, el valor promedio de compra de estos consumidores es bastante alto con un valor 11406 euros aproximademente. Por lo que se identifica a este segmento como uno de los más rentables.