Certificado Profesional de Análisis de Datos de Google

Ing. Ignacio Jiménez Mota, Analista de Datos. Contacto: ignacio.jmota@gmail.com

Ciudad de México, México. 31 de marzo del 2023

Trabajo final: Resolución del caso práctico número 1

"Análisis de las tendencias de ventas en tres sucursales de un supermercado usando lenguaje de programación R"

Introducción Este documento es suplementario al principal (Proyecto1_Programación_R.pdf) que se puede encontar tanto en el portafolio de Google Drive como en el de GitHub. Es recomendable leer ese documento primero y después continuar con este, cuyo objetivo principal es mostrar el código utilizado para responder las preguntas empresariales y las resultados del análisis. La fuente de la base de datos se encuentra en Kaggle.

Preparar Los siguientes paquetes se instalaron y cargaron para ser usados en este proyecto:

```
library("here")
library("skimr")
library("janitor")
library("readr")
library("dplyr")
library("tidyr")
library("ggplot2")
```

Después, se importó la base de datos como *data frame* y se le asignó un nombre:

```
supermarket_sales_df1 <- read_csv("supermarket_sales.csv")
```

Con la función 'skim_without_charts()' se puede obtener un resumen del conjunto de datos: número de filas y columnas,tipo de datos en las columnas, valores únicos por columna, celdas vacías, o celdas con valores que incluyen espacios.

```
skim_without_charts(supermarket_sales_df1)
```

Table 1: Data summary

Name	supermarket_sales_df1
Number of rows	1000
Number of columns	17
Column type frequency:	
character	8
difftime	1
numeric	8
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Invoice ID	0	1	11	11	0	1000	0
Branch	0	1	1	1	0	3	0
City	0	1	6	9	0	3	0
Customer type	0	1	6	6	0	2	0
Gender	0	1	4	6	0	2	0
Product line	0	1	17	22	0	6	0
Date	0	1	8	9	0	89	0
Payment	0	1	4	11	0	3	0

Variable type: difftime

skim_variable	n_missing	complete_rate	min	max	median	n_unique
Time	0	1	36000 secs	75540 secs	55140 secs	506

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Unit price	0	1	55.67	26.49	10.08	32.88	55.23	77.94	99.96
Quantity	0	1	5.51	2.92	1.00	3.00	5.00	8.00	10.00
Tax 5%	0	1	15.38	11.71	0.51	5.92	12.09	22.45	49.65
Total	0	1	322.97	245.89	10.68	124.42	253.85	471.35	1042.65
cogs	0	1	307.59	234.18	10.17	118.50	241.76	448.91	993.00
gross margin percentage	0	1	4.76	0.00	4.76	4.76	4.76	4.76	4.76
gross income	0	1	15.38	11.71	0.51	5.92	12.09	22.45	49.65
Rating	0	1	6.97	1.72	4.00	5.50	7.00	8.50	10.00

A partir de la vista previa anterior, se observa que los nombres de las columnas podrían mejorarse para facilitar su análsis y evitar errores en el futuro. La función 'clean_names()' garantiza que solamente existan caracteres, números y guiones bajos en las nombres de las columnas. De este modo,se creó un nuevo *data frame* para guardar los cambios.

```
supermarket_sales_df2 <- clean_names(supermarket_sales_df1)
```

Con la función 'colnames()'se comprobó que se hayan guardado los cambios en el nuevo *data frame* (df).

```
colnames(supermarket_sales_df2)
```

```
##  [1] "invoice_id"             "branch"
##  [3] "city"                   "customer_type"
##  [5] "gender"                 "product_line"
##  [7] "unit_price"             "quantity"
##  [9] "tax_5_percent"          "total"
## [11] "date"                   "time"
## [13] "payment"                "cogs"
## [15] "gross_margin_percentage" "gross_income"
## [17] "rating"
```

Analizar A continuación, se responderán las preguntas empresariales y se explicarán algunos detalles. Debido a la naturaleza de todas las preguntas, salvo la número 5, será más eficiente contestarlas con una visualización de datos donde se categorice según el grupo que indique la pregunta y, en forma de barras, se representará la distribución correspondiente.

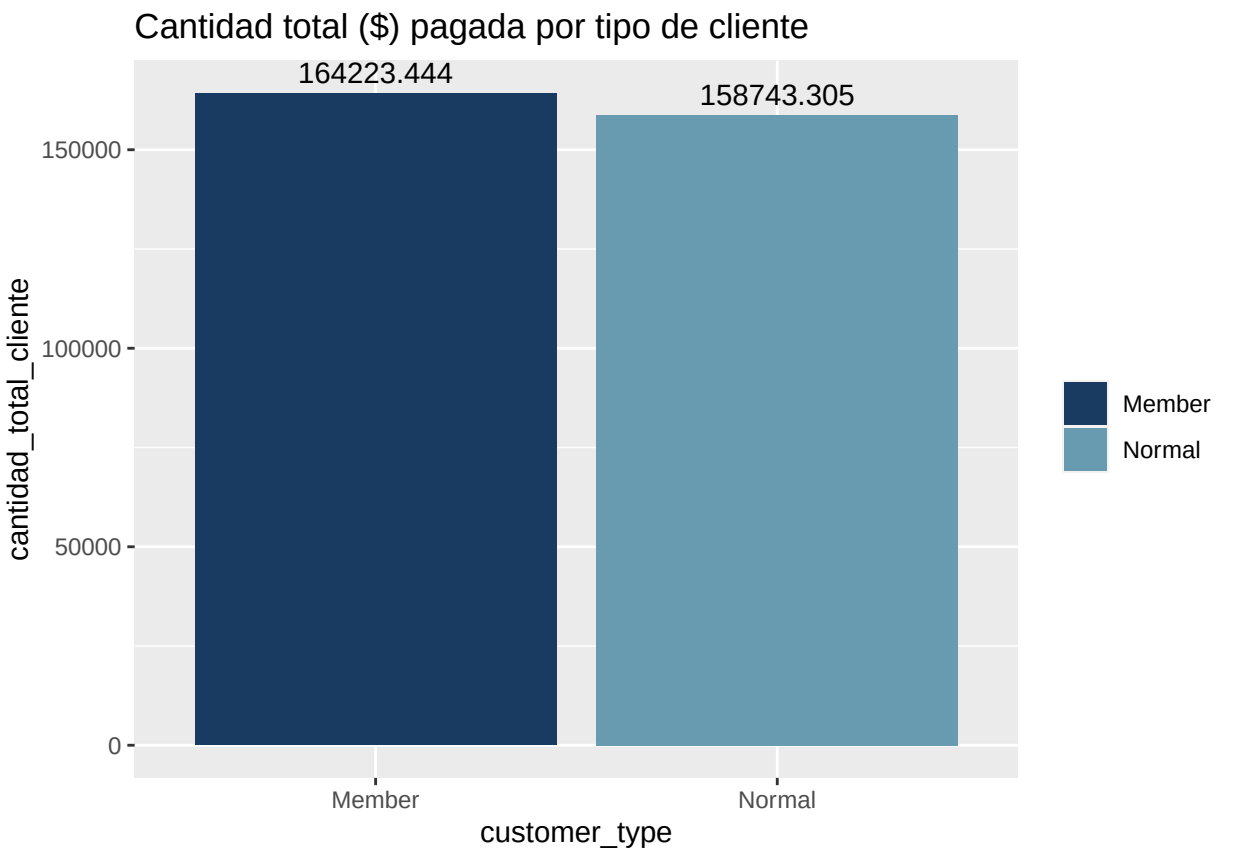
1. ¿Qué tipo de cliente acumula la mayor cantidad total pagada por sus compras?

Se creó un nuevo df donde se agrupe por tipo de cliente y se sume el valor de la columna **total** para obtener la cantidad total pagada al supermercado por tipo de cliente. **df1** contiene dicho valor para cada tipo de cliente:

```
df1 <- supermarket_sales_df2 %>%
  group_by(customer_type) %>%
  summarise(cantidad_total_cliente = sum(total))
```

Una vez preparado el nuevo df, se representó en un gráfica a través de ggplot, incluyendo algunas capas para el título y las etiquetas:

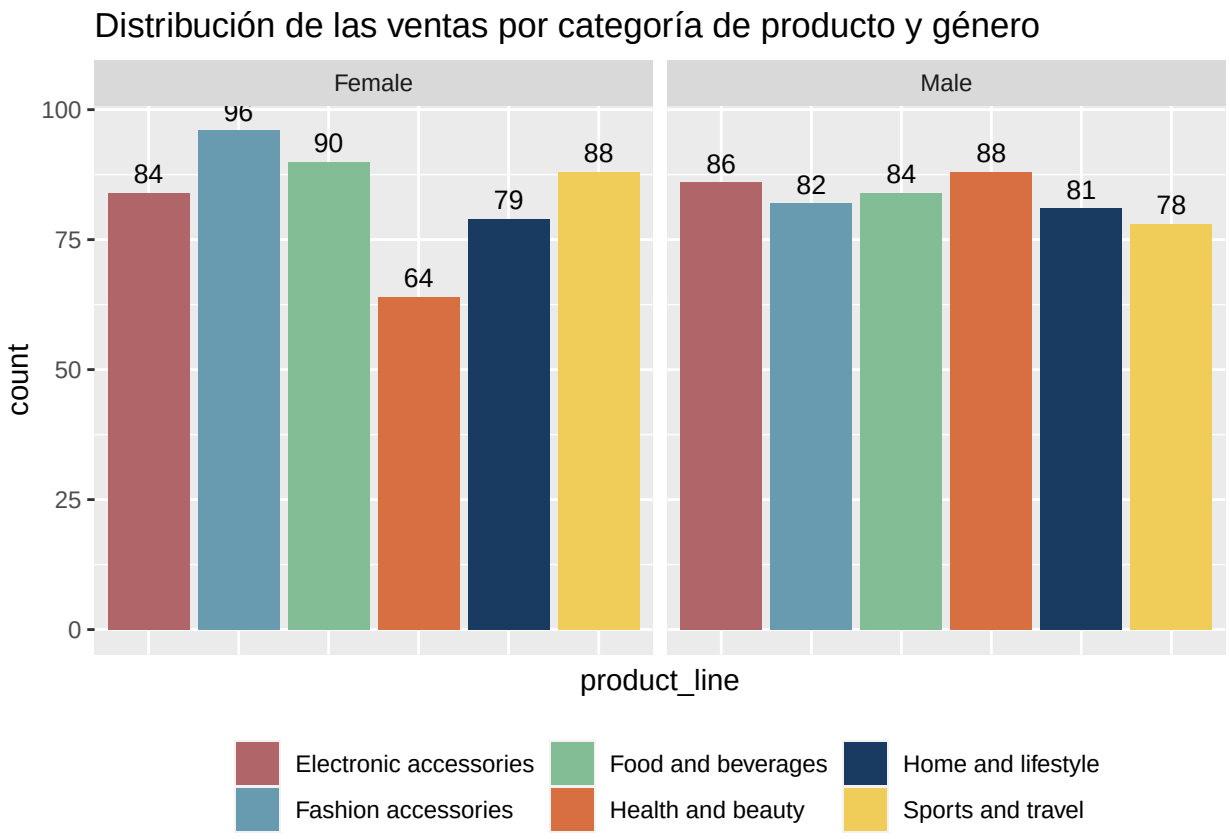
```
ggplot(df1,
  aes(x = customer_type, y = cantidad_total_cliente, fill = customer_type)) +
  geom_bar(stat = "identity") +
  labs(title = "Cantidad total ($) pagada por tipo de cliente") +
  geom_text(aes(label = cantidad_total_cliente), vjust = -0.5) +
  scale_fill_manual(values=c("#1A3B61", "#689BAF")) +
  theme(legend.title = element_blank())
```



En dicha visualización se observa que el grupo de clientes que acumula una mayor cantidad de dinero por sus compras, aunque no por mucho, es *Member* con $ 164,223.444 (50. 85 % de la cantidad total paga al supermercado); mientras que el grupo *Normal* acumula $ 158,743.305 (49.15 %).

2. ¿Cuál es la categoría de producto que con más frecuencia consumen las mujeres y cuál los hombres?

```
ggplot(data = supermarket_sales_df2,
  aes(x = product_line, fill = product_line)) + geom_bar() +
  facet_wrap(~gender) +
  labs(title =
         "Distribución de las ventas por categoría de producto y género") +
  theme(legend.position = "bottom", axis.text.x = element_blank(),
        axis.ticks.x = element_blank(), legend.title = element_blank()) +
  geom_text(aes(label = after_stat(count)), stat = 'count',
            position = position_dodge(0.9), vjust = -0.5, size = 3.5) +
  scale_fill_manual(values=c("#B06669", "#689BAF", "#82BD95", "#D86F41",
                             "#1A3B61", "#F0CD59"))
```

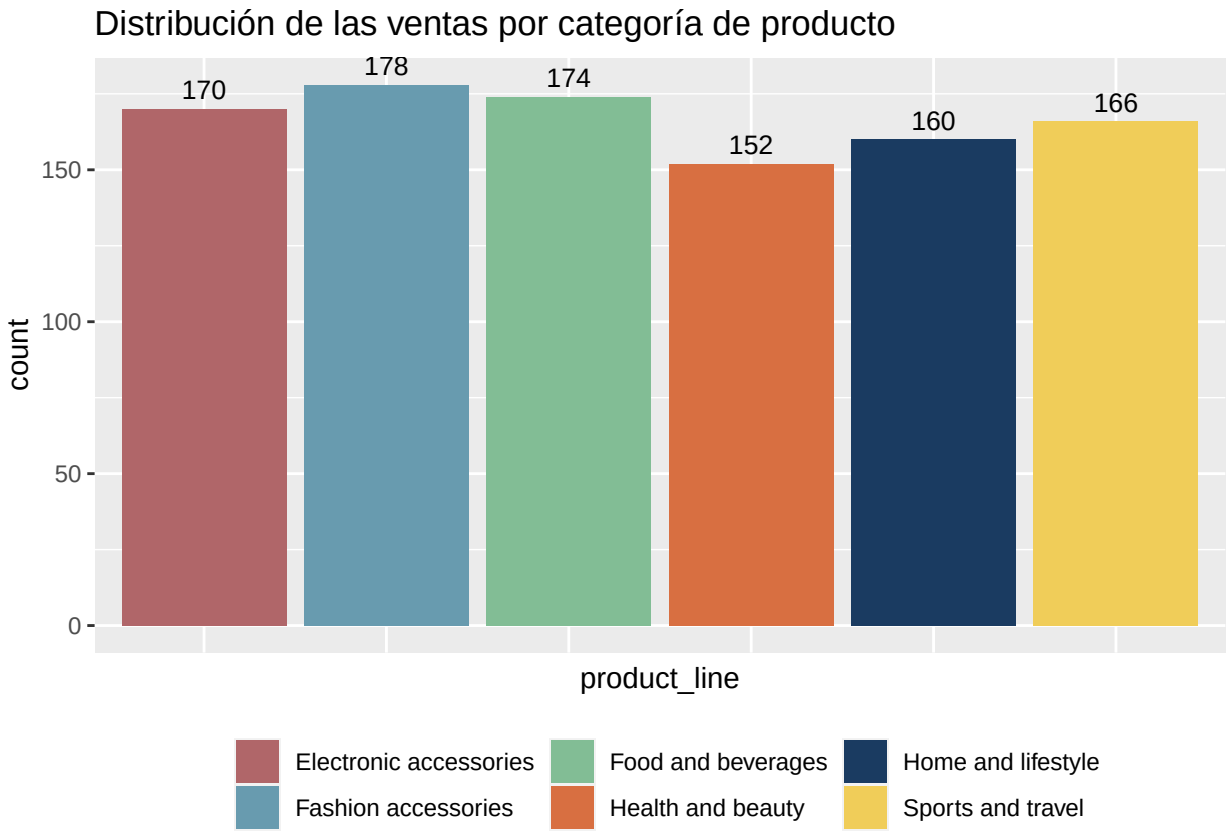


En dicha visualización se observa que el grupo de clientes de género femenino consume con mayor frecuencia la categoría *Fashion accessories* (96 registros) y con menor frecuencia, *Healh and beauty* (64 registros); por otra parte, el grupo de clientes del género masculino consume con mayor frecuencia la categoría *Healh and beauty* (88 registros) y con menor frecuencia, *Sports and travel* (78 registros). Aunque, en ambos grupos, no existe demasiada diferencia entre las categorías.

3. ¿Cuál es la distribución de las ventas por categoría de producto?

```
ggplot(data = supermarket_sales_df2,
  aes(x = product_line, fill = product_line)) + geom_bar() +
  labs(title = "Distribución de las ventas por categoría de producto") +
  theme(legend.position = "bottom", axis.text.x = element_blank(),
        axis.ticks.x = element_blank(), legend.title = element_blank()) +
```

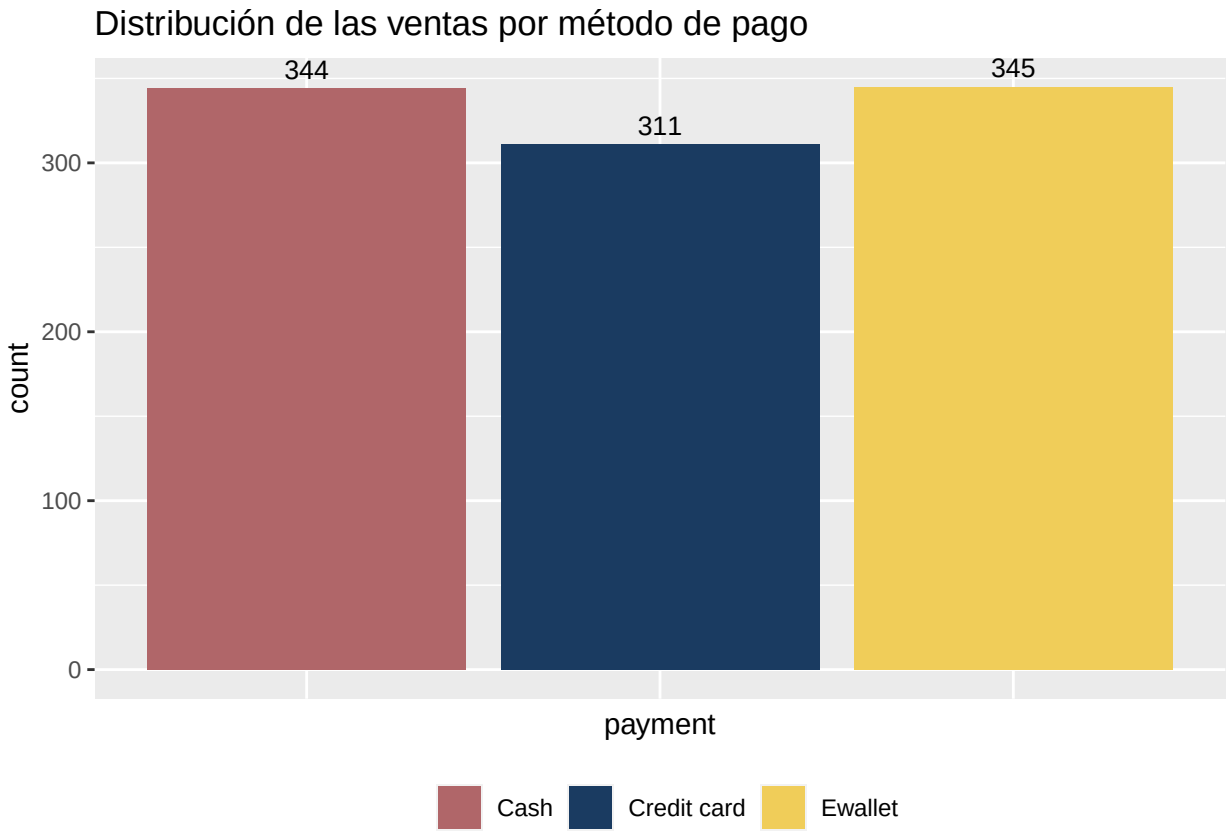
```
  geom_text(aes(label = after_stat(count)), stat = 'count',
            position = position_dodge(0.9), vjust = -0.5, size = 3.5) +
  scale_fill_manual(values=c("#B06669", "#689BAF", "#82BD95", "#D86F41",
                             "#1A3B61", "#F0CD59"))
```



En dicha visualización se observa que el grupo *Fashion accessories* es el que mayor frecuencia de ventas representa, 178 registros; por otra parte, *Healh and beauty* es el grupo con menor frecuencia de ventas, 152 registros. Aunque, en los seis grupos, no existe demasiada diferencia entre la frecuencia de ventas.

4. ¿Cuál es la distribución de las ventas por método de pago?

```
ggplot(data = supermarket_sales_df2,
  aes(x = payment, fill = payment)) + geom_bar() +
  labs(title = "Distribución de las ventas por método de pago") +
  theme(legend.position = "bottom", axis.text.x = element_blank(),
        axis.ticks.x = element_blank(), legend.title = element_blank()) +
  geom_text(aes(label = after_stat(count)), stat = 'count',
            position = position_dodge(0.9), vjust = -0.5, size = 3.5) +
  scale_fill_manual(values=c("#B06669","#1A3B61", "#F0CD59"))
```



En dicha visualización se observa que el grupo *Ewallet* es el que mayor frecuencia de ventas representa, 345 registros; por otra parte, *Credit card* es el grupo con menor frecuencia de ventas, 311 registros. Aunque, en los tres grupos, no existe demasiada diferencia entre la frecuencia de ventas, pues en segundo lugar se encuentra el grupo *Cash*, con 344 registros.

5. ¿Cuál es el ingreso bruto total del supermercado?

```
ingreso_bruto_total <- supermarket_sales_df2 %>%
  summarise(ingreso_bruto = sum(gross_income))
ingreso_bruto_total
```

```
## # A tibble: 1 x 1
##   ingreso_bruto
##           <dbl>
## 1        15379.
```

Para este caso, bastó con realizar un *pipe* para crear la variable **ingreso_bruto_total** y sumar los valores de la columna **gross_income**, el resultado es el ingreso bruto total del supermercado (considerando las tres sucursales): $ 15,379.00. Los valores de la columna **gross_income** corresponden a la resta entre la columna **total** y **cogs**.

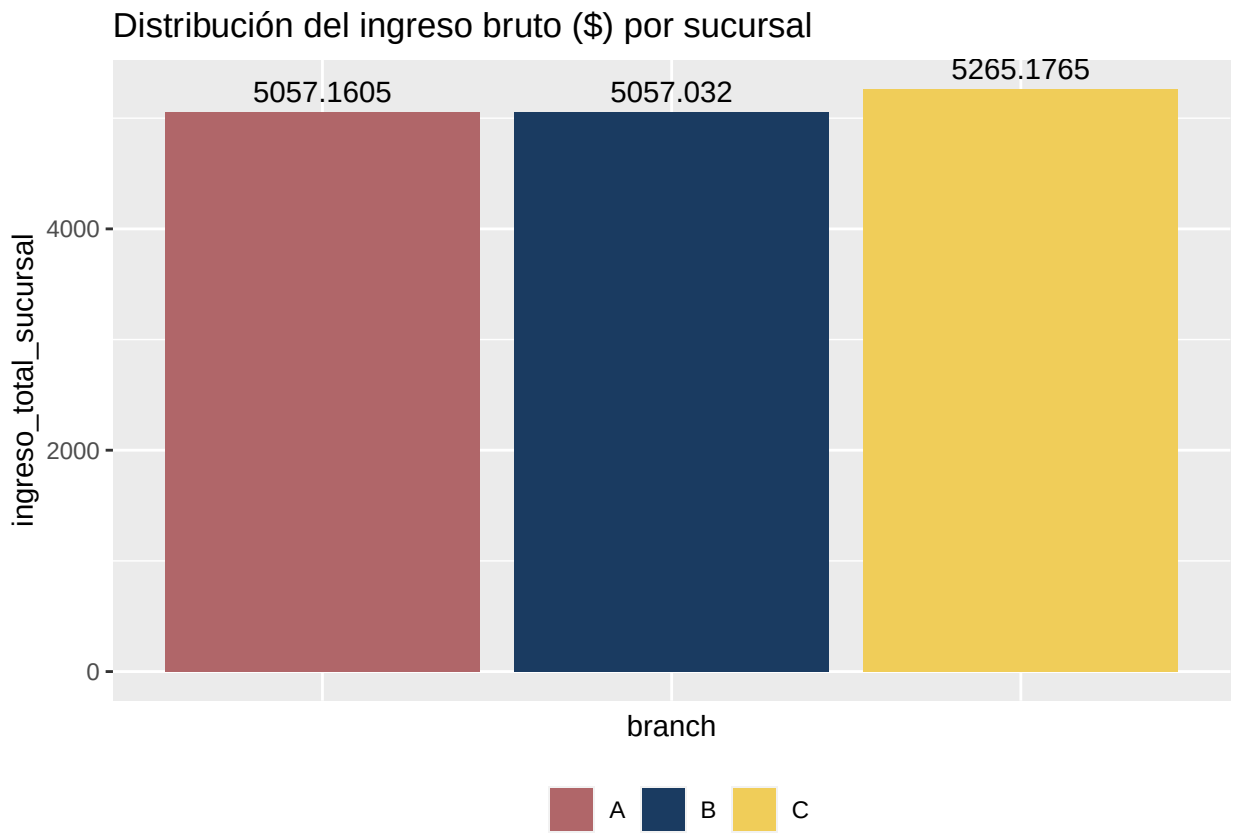
6. ¿Cuál es la distribución de ingresos brutos por sucursal?

Se creó un nuevo df donde se agrupe por sucursal y se sume el valor de la columna **gross_income** para obtener el ingreso bruto distribuido por sucursal. **df2** contiene dicho valor para cada sucursal:

```
df2 <- supermarket_sales_df2 %>%
  group_by(branch) %>%
  summarise(ingreso_total_sucursal = sum(gross_income))
```

Una vez preparado el nuevo df, se representó en un gráfica a través de ggplot, incluyendo algunas capas para el título y las etiquetas:

```
ggplot(df2,
  aes(x = branch, y = ingreso_total_sucursal, fill = branch)) +
  geom_bar(stat = "identity") +
  labs(title = "Distribución del ingreso bruto ($) por sucursal") +
  geom_text(aes(label = ingreso_total_sucursal), vjust = -0.45) +
  theme(legend.position = "bottom", axis.text.x = element_blank(),
        axis.ticks.x = element_blank(), legend.title = element_blank()) +
  scale_fill_manual(values=c("#B06669","#1A3B61", "#F0CD59"))
```



En dicha visualización se observa que la sucursal **C** es la que mayor ingreso bruto representa para el supermercado, aportando $ 5,265.1765 (34.24 % del ingreso bruto total del supermercado); por otra parte, las sucursales **A** y **B** están prácticamente empatadas, con ingresos brutos de $ 5,057.1605 (32.88 %) y $ 5057.032 (32.88 %), respectivamente. Aunque, como en los casos anteriores, no existe demasiada diferencia entre la distribución de ingresos brutos por sucursal.

7. ¿Qué sucursal tiene la mejor calificación promedio por lo clientes?

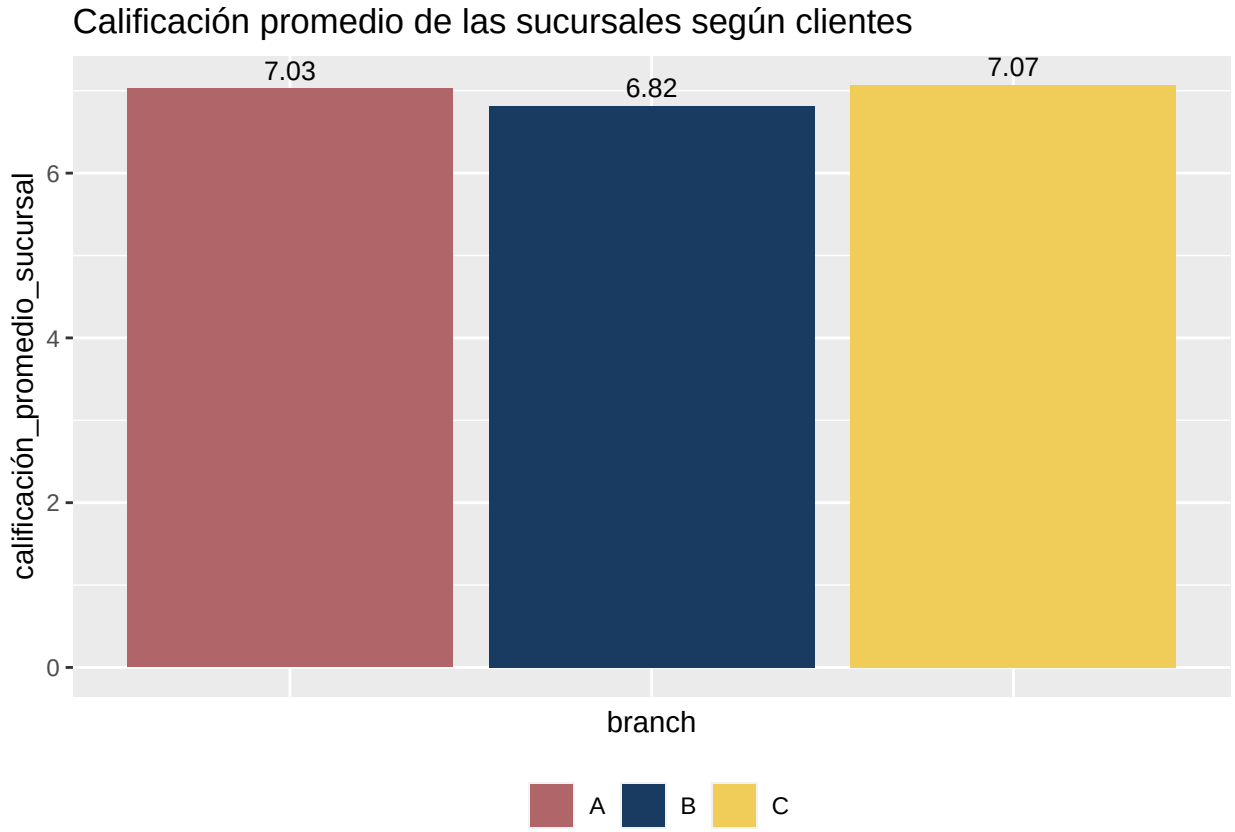
Se creó un nuevo df donde se agrupe por sucursal y se obtenga el valor promedio de la columna **rating** para obtener la calificación promedio de cada sucursal otorgada por los clientes. **df3** contiene dicho valor para cada sucursal:

```
df3 <- supermarket_sales_df2 %>%
  group_by(branch) %>%
  summarise(calificación_promedio_sucursal = mean(rating))
```

Una vez preparado el nuevo df, se representó en un gráfica a través de ggplot, incluyendo algunas capas para

el título y las etiquetas:

```
ggplot(df3,
  aes(x = branch, y = calificación_promedio_sucursal, fill = branch)) +
  geom_bar(stat = "identity") +
  labs(title = "Calificación promedio de las sucursales según clientes") +
  geom_text(aes(label =
          sprintf("%0.2f", round(calificación_promedio_sucursal, digits = 2))),
          position = position_dodge(0.9), vjust = -0.45, size = 3.5) +
  theme(legend.position = "bottom", axis.text.x = element_blank(),
        axis.ticks.x = element_blank(), legend.title = element_blank()) +
  scale_fill_manual(values=c("#B06669","#1A3B61", "#F0CD59"))
```



En dicha visualización se observa que las sucursales **A** y **C** son las que cuentan con mejor calificación promedio, con valores 7.03 y 7.07, respectivamente; por otra parte, la sucursal **B**, es la sucursal con menor calificación promedio, 6.82. Aunque, en las tres sucursales no existe demasiada diferencia entre las calificaciones que han recibido por parte de los clientes.

Actuar A través del análisis realizado a los registros de ventas de un supermercado en tres de sus sucursales, se observaron algunas tendencias: el tipo de cliente que más ingresos representa al supermercado es *Member*; aunque, como se verá en los resultados siguientes, no hay demasiada diferencia entre las categorías; para este mismo caso, sería considerable una campaña de marketing para consolidar como *Member* a los clientes que todavía no lo son y que en este momento representan el 49.14 % de las ventas.

En cuanto a la distribución de las ventas por género de cliente y por categoría de producto, se observa bastante equilibro entre cada una; quizá, un caso a considerar para mejorarlo es la categoría de producto *Health and beauty*: la de mayores ventas para el género *Male*, sin embargo, la de menores ventas para el

género *Female*. Sería conveniente averiguar la razón por la cuál una categoría tan popular para un género, es tan impopular para el otro; lo anterior, averiguando qué productos en específico representan las mayores ventas por género y, de este modo, establecer una estrategia también más específica. Esta misma categoría, *Health and beauty*, en la distribución general de las ventas, es decir, considerando ambos géneros, es la que menores ventas representa: la estrategia mencionada anteriormente también contribuiría en este caso.

Respecto a la distribución de las ventas por método de pago, se observó que la categoría *Ewallet* representa la mayoría en preferencia por los clientes; si bien es cierto que no por mucho, pues le sigue de cerca la categoría *Cash*. Aquí también sería relevante una campaña informativa donde se le explique a los clientes las ventajas de usar la forma de pago *Ewallet* en lugar de *Cash*; de este modo, las sucursales tendrían una reducción de costos respecto a la logística y traslado del efectivo a los respectivos bancos.

El ingreso bruto total del supermercado (considerando los tres meses de registro y a las tres sucursales) fue de $ 15,379.00. Y la distribución de este por sucursal fue de la siguiente manera, el 34.24 % lo aportó la sucursal **C** (siendo la de mayor aportación), mientras que las sucursales **A** y **B** están prácticamente empatadas en este ámbito, con el 32.88 % cada una. Una estrategia para mejorar el rendimiento de las tres sucursales estaría basada en los resultados de la pregunta 7: ¿qué sucursal posee la mejor calificación promedio otorgada por lo clientes? Curiosamente o, mejor dicho, por alguna razón en particular, la sucursal mejor calificada es la que más ingresos representa para el supermercado: la **C**. La estrategia que se propone estaría fundamentada en un análisis específico par conocer la razón por la cuál tiene una mejor calificación y un mejor rendimiento. Sería necesario recolectar datos a través de encuestas o a través de redes sociales, inclusive, de sitios como Google Maps, donde las personas pueden plasmar su opinión sobre los lugares que visitan. Conociendo dichas razones fundamentales, la opinión del público, por ejemplo, se podría mejorar una parte del rendimiento de las sucursales desde una perspectiva de atención al cliente.

Ing. Ignacio Jiménez Mota, Analista de Datos

Contacto: ignacio.jmota@gmail.com