Certificado Profesional de Análisis de Datos de Google

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”

Presenta:

Ing. Ignacio Jiménez Mota, Analista de Datos

Ciudad de México, México

31 de marzo del 2023

ignacio.jmota@gmail.com

Contenido

Introducción

El presente documento tiene la final de documentar todo el proceso de análisis de la base de datos seleccionada, así como la descripción de la misma, tanto en sus características como en su origen. De igual forma, el objetivo general es resolver las preguntas empresariales que guían este proyecto y con las cuales se planea descubrir las tendencias mencionadas en el título del trabajo.

Respecto a la base de datos, es una colección de registros de ventas recopilados durante tres meses y en tres sucursales de un supermercado. El conjunto está estructurado a partir de una clave primara, Invoice ID (número de identificación de la factura) y, cada observación consta de campos como Branch (sucursal donde se llevó a cabo la venta), City (la ciudad donde se ubica la sucursal), Customer type (el tipo de cliente, normal o asociado), Gender (el género del cliente), Product line (la categoría del producto vendido), Unit price (el precio unitario de artículo comprado por el cliente), Quantity (la cantidad de productos comprados por el cliente), Tax 5 % (el impuesto asociado a la compra del cliente), Total (la cantidad total pagada por el cliente en la respectiva factura), Date (fecha de la venta), Time (hora de la venta), Payment (método de pago), COGS (Cost of Goods Sold, costo de los bienes vendidos), Gross Margin Percentage (porcentaje del margen de ingreso bruto), Gross Income (ingreso bruto) y Rating (calificación del cliente en su experiencia de compra en una escala del 1 al 10).

El conjunto de datos fue extraído de Kaggle. La persona que lo publicó indica que pude ser usado con propósitos de análisis predictivo; asimismo, indica que el propietario es Aung Pyae, estudiante de M.Sc en Data Science y Business Analytics. Es necesario aclarar que el autor únicamente indica que los precios de los productos están dados en “$”; para facilitar el análisis, se asumirá que son dólares estadounidenses (USD).

Escenario

Con el perfil y desde la perspectiva de analista de datos, se conducirá este proyecto que involucrará actividades desde la definición de la tarea empresarial (o las preguntas empresariales), el recorrido, tanto de preparación como de análisis, y hasta la presentación de las recomendaciones basadas en datos. El tema que se ha elegido corresponde al registro de ventas en un supermercado. Con base en dicho registro, se harán las preguntas empresariales, se garantizará la integridad de los datos, se realizará el análisis y se crearán las visualizaciones necesarias para presentar a los *stakeholders* a través de una presentación elaborada a través de la herramienta R Markdown.

Preguntar

¿Cuál es la tarea empresarial que se desea resolver?

Se desean conocer diversas tendencias en las ventas de tres sucursales de un supermercado a través del lenguaje de programación R.

¿Qué métricas en específico se planea conocer?

1. ¿Qué tipo de cliente acumula la mayor cantidad total pagada por sus compras?
2. ¿Cuál es la categoría de producto que con más frecuencia consumen las mujeres y cuál los hombres?
3. ¿Cuál es la distribución de las ventas por categoría de producto?
4. ¿Cuál es la distribución de las ventas por método de pago?
5. ¿Cuál es el ingreso bruto total del supermercado?
6. ¿Cuál es la distribución de ingresos brutos por sucursal?
7. ¿Qué sucursal tiene la mejor calificación promedio por lo clientes?

¿Quiénes son los *stakeholders*?

Para este proyecto, son los líderes de cada sucursal, así como los líderes del centro de distribución que las provee.

Preparar

El conjunto de datos seleccionado para este proyecto, se nombró "supermarket_sales" y es un archivo CSV. Está guardado en el *working directory* establecido previamente para trabajar con R y cuya dirección es "C:/Users/Ignacio Jiménez Mota/Documents/RStudio/R_Projects/3_CertificadoProfesionalAnalisisDatosGoogle_Proyecto1/Analysis-of-supermarket-sales-using-R". La base de datos se mantendrá íntegra, es decir, no se le modificará en su contenido, simplemente se limpiará con herramientas de R de tal manera que no afecten los datos originales.

En el mismo *working directory* se crearon dos archivos más: "CPADG_Proyecto1_Programación_R", un *script* con extensión R en donde quedará registrado todo el proceso de análisis a través de los códigos necesarios para ello (importación de datos, limpieza, creación de visualizaciones, comentarios, etc.); también "CPADG_Proyecto1_Programación_R", un archivo en formato R Markdown (extensión RMD) con la finalidad de exportar un reporte más formal (la presentación antes mencionada, en HTML y PDF), donde se pueda apreciar el código que se produjo para este análisis así como los resultados respectivos.

Procesar

En el *script*, "CPADG_Proyecto1_Programación_R" con extensión R, han quedado registrados los códigos necesarios para instalar y cargar los paquetes que se utilizarán en este paquete: 4 de los 8 que conforman el *tidyverse* (*readr, dplyr, tidyr, ggplot2*); *here*, que facilita la consulta de los datos; *skimr*, que facilita el resumen de los datos; *janitor*, que contiene funciones de limpieza.

Documentación de la manipulación y limpieza de los datos

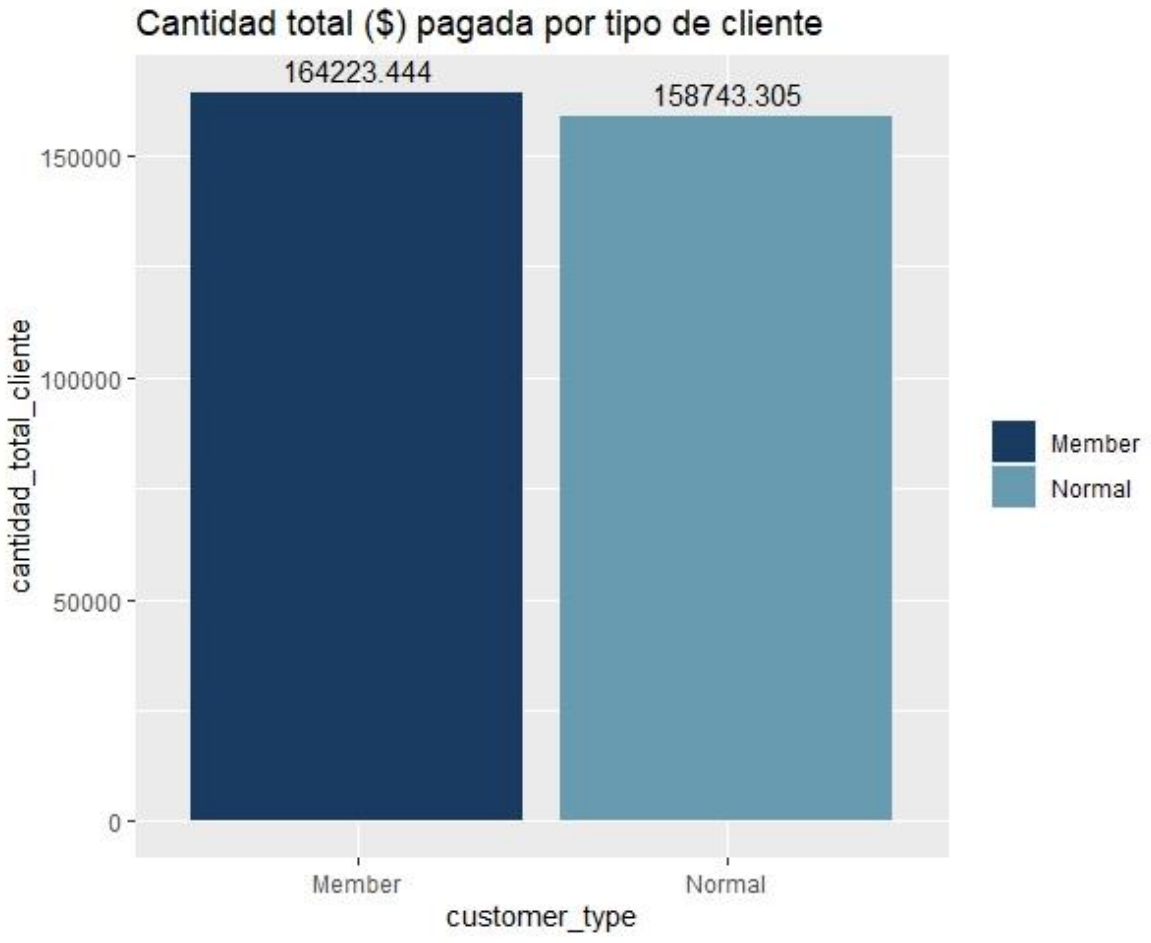
1. Uso de la función read_csv() para importar la base de datos y asignarle un nombre como *data frame* (df): “supermarket_sales_df1”.
2. Uso de algunas funciones para obtener una vista previa del conjunto de datos:
    - La función skim_without_charts se usó para obtener un resumen de las características de las observaciones y los campos; por ejemplo, hay 1000 filas y 17 columnas; estas últimas contienen 3 tipos de datos: *character* (8 columnas), *difftime* (1 columna), *numeric* (8 columnas). Además, a través de los resultados de esta función, se puede constatar que la columna Invoice ID, la clave primaria, contiene 1000 observaciones únicas, es decir, que no existen registros duplicados; también, que los demás registros no incluyen celdas vacías ni valores con espacios.
    - La función glimpse() muestra la cantidad de columnas y filas, así como el nombre de cada columna, el tipo de datos que contiene y algunos ejemplos de ellos entre comillas.
    - La función head() muestra un tibble de 6 X 17 (las seis primeras filas y las 17 columnas que componen el conjunto de datos).
3. A partir de las vistas previas anteriores, se observa que los nombres de las columnas podrían mejorarse para facilitar su análisis y evitar errores en el futuro:
    - La función clean_names() se utilizó para garantizar que solamente existan caracteres, números y guiones bajos en las nombres de las columnas. Así mismo, se creó un nuevo df para guardar los cambios: “supermarket_sales_df2”.
    - La función colnames() se utilizó para comprobar que se hayan guardado los cambios en el nuevo df.
4. Con lo anterior se garantiza la limpieza, integridad y disposición de los datos para ser analizados.

Analizar

A continuación, se responderán las preguntas empresariales y se indicarán las herramientas utilizadas:

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

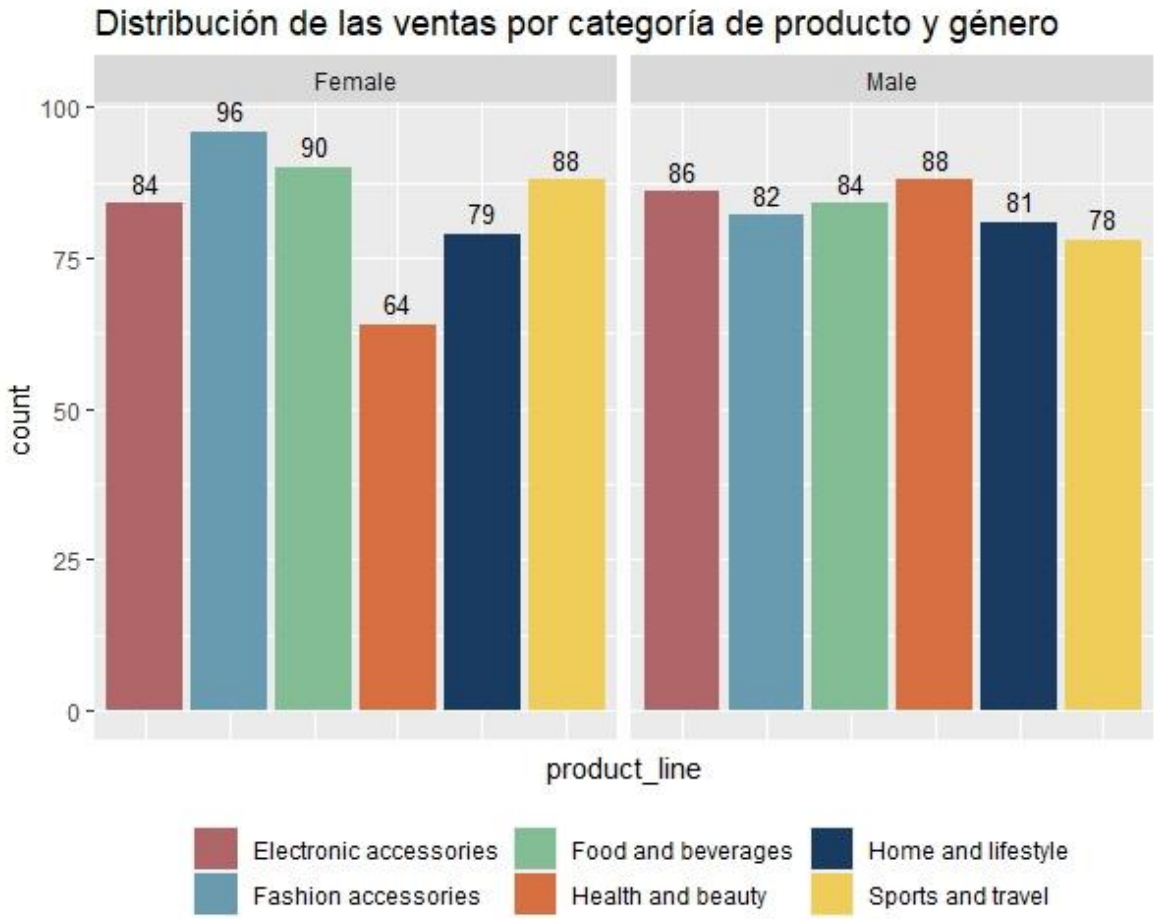
Debido a la naturaleza de esta pregunta, será más eficiente contestarla con una visualización donde se categorice al cliente y, en forma de barras, se represente la cantidad total pagada por sus compras. Para lo anterior se utilizará la función ggplot() y las capas subsecuentes quedarán registradas en el *scrip* antes mencionado, "CPADG_Proyecto1_Programación_R". La visualización resultante se guardará en el *working directory* actual bajo el nombre "Dataviz_Pregunta1_CPADG_Proyecto1".



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?**

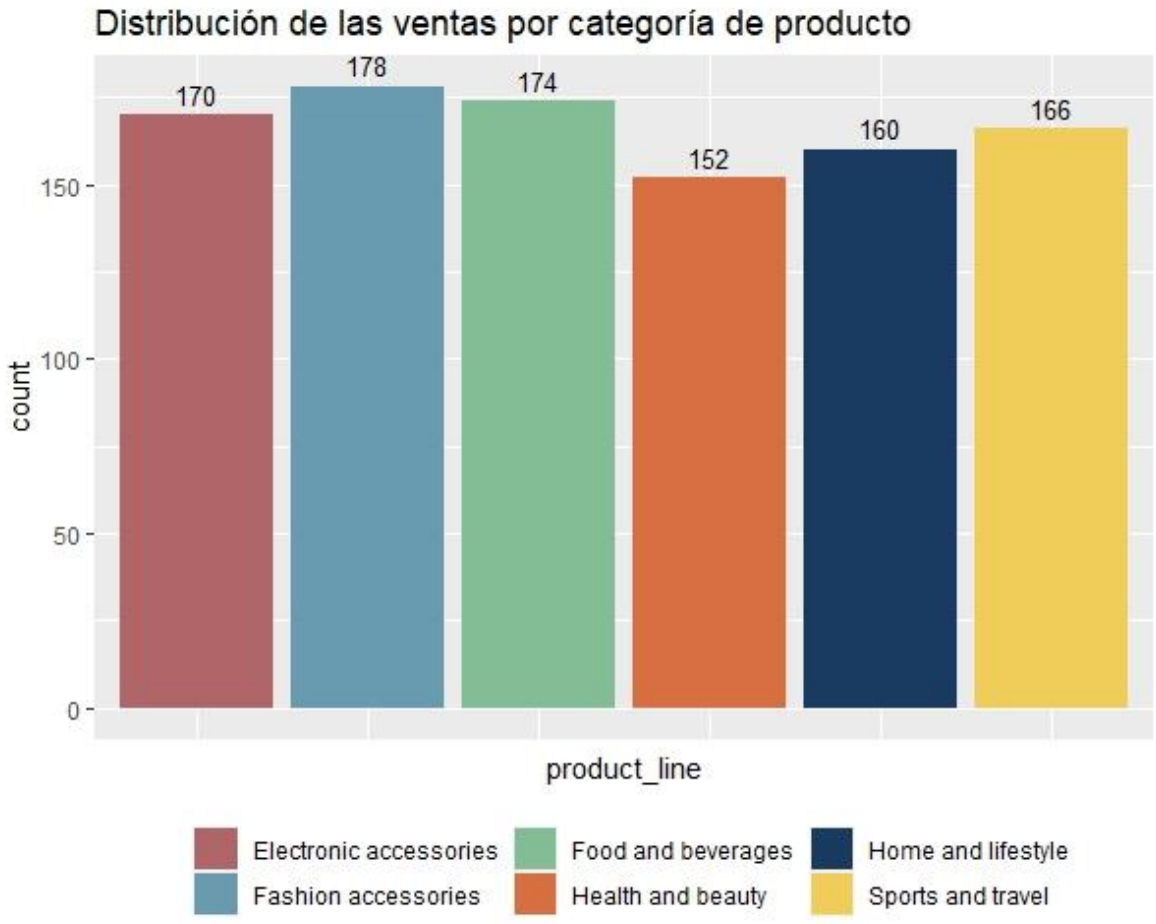
También será más eficiente contestarla con una visualización donde se categorice por género y, en forma de barras, se represente la distribución por categoría de producto. Para lo anterior se utilizará la función ggplot y las capas subsecuentes quedarán registradas en el *scrip* antes mencionado, "CPADG_Proyecto1_Programación_R". La visualización resultante se guardará en el *working directory* actual bajo el nombre "Dataviz_Pregunta2_CPADG_Proyecto1".



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?

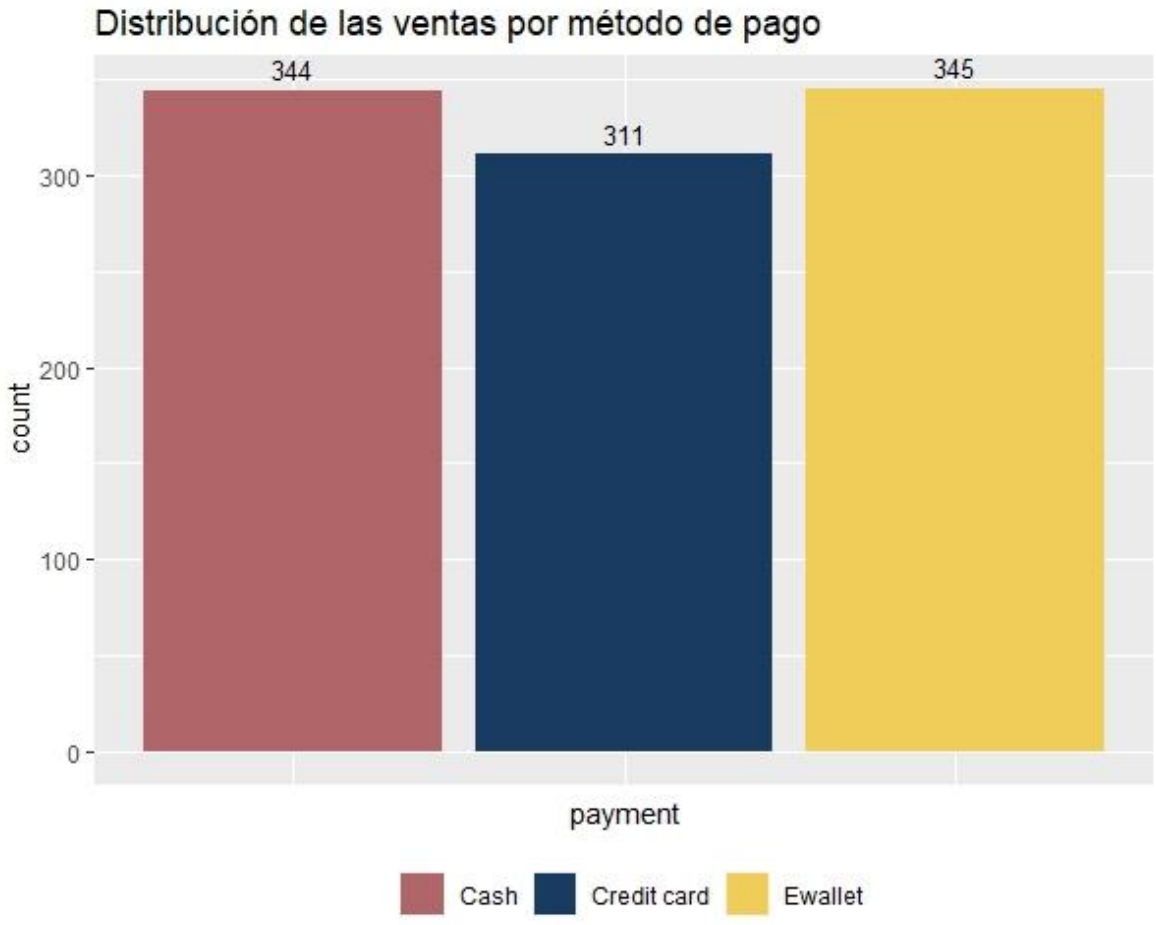
También será más eficiente contestarla con una visualización donde se categorice por tipo de género y, en forma de barras, se represente la distribución de las ventas. Para lo anterior se utilizará la función ggplot y las capas subsecuentes quedarán registradas en el scrip antes mencionado, "CPADG_Proyecto1_Programación_R". La visualización resultante se guardará en el working directory actual bajo el nombre "Dataviz_Pregunta3_CPADG_Proyecto1".



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?

También será más eficiente contestarla con una visualización donde se categorice por tipo de método de pago y, en forma de barras, se represente la distribución de las ventas. Para lo anterior se utilizará la función ggplot y las capas subsecuentes quedarán registradas en el *scrip* antes mencionado, "CPADG_Proyecto1_Programación_R". La visualización resultante se guardará en el *working directory* actual bajo el nombre "Dataviz_Pregunta4_CPADG_Proyecto1".



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?**

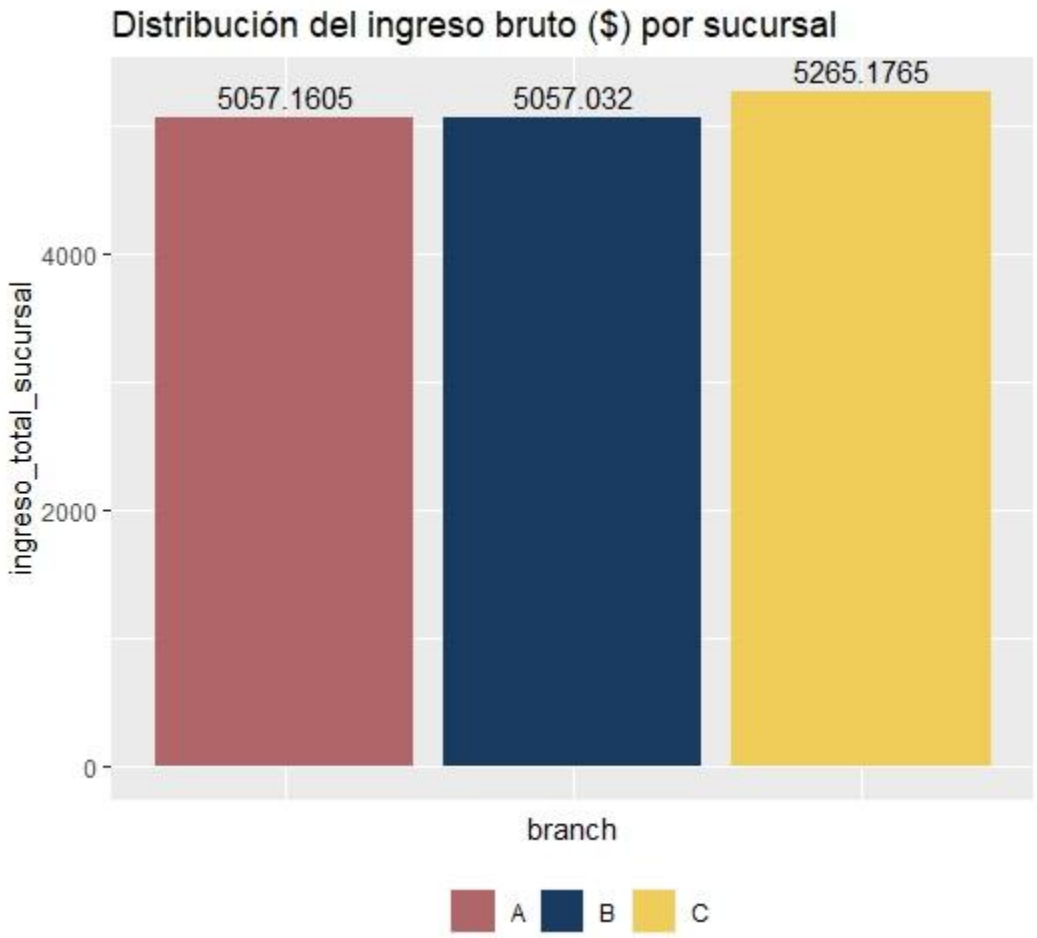
Para este caso, basta con realizar un *pipe* para crear una variable y sumar los valores de la columna "gross_income", el resultado es el ingreso bruto total del supermercado (considerando las tres sucursales): $ 15,379.00.

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?**

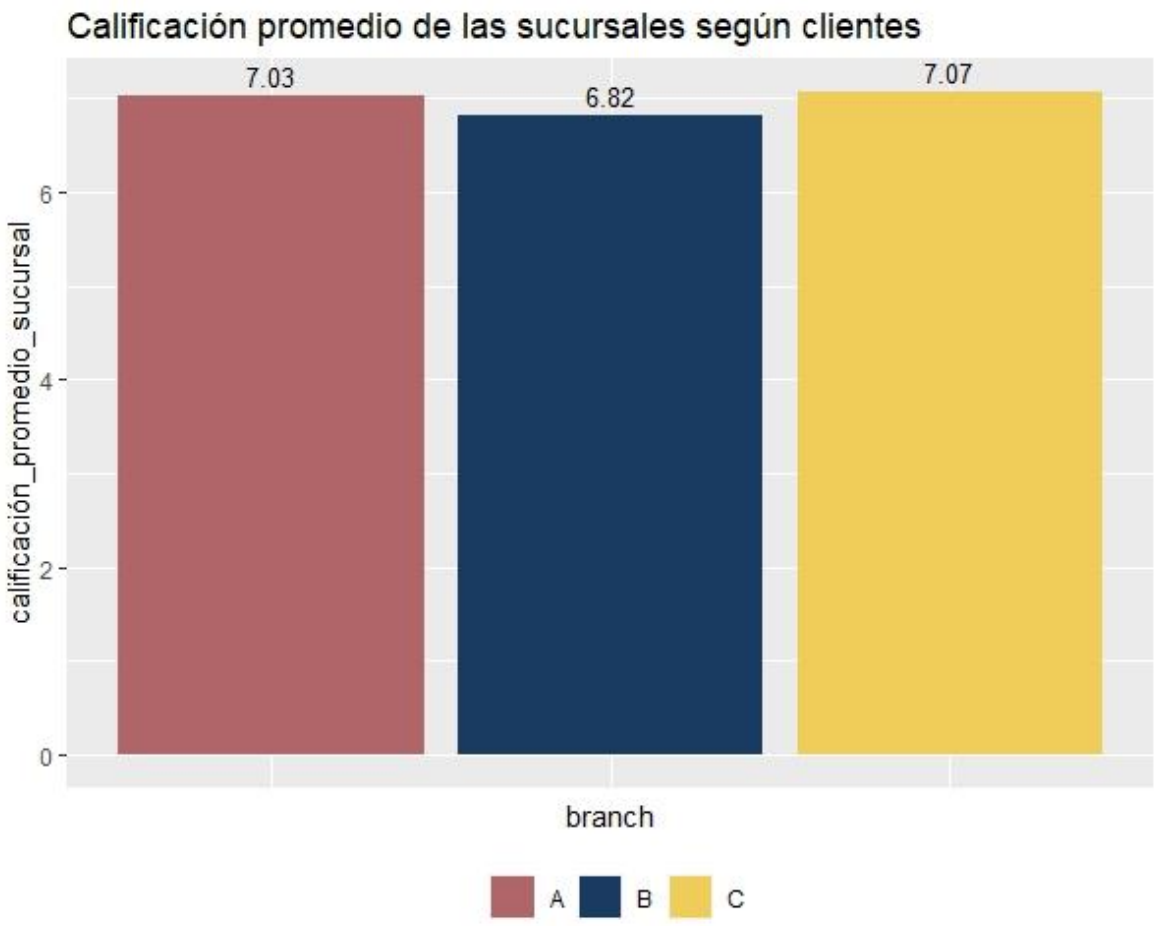
También será más eficiente contestarla con una visualización donde se categorice por sucursal y, en forma de barras, se represente la distribución del ingreso bruto del supermercado. Para lo anterior se utilizará la función ggplot y las capas subsecuentes quedarán registradas en el *scrip* antes mencionado, "CPADG_Proyecto1_Programación_R". La visualización resultante se guardará en el *working directory* actual bajo el nombre "Dataviz_Pregunta6_CPADG_Proyecto1".

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?

También será más eficiente contestarla con una visualización donde se categorice por sucursal y, en forma de barras, se represente la calificación promedio otorgada por los clientes. Para lo anterior se utilizará la función ggplot y las capas subsecuentes quedarán registradas en el *scrip* antes mencionado, "CPADG_Proyecto1_Programación_R". La visualización resultante se guardará en el *working directory* actual bajo el nombre "Dataviz_Pregunta7_CPADG_Proyecto1".



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.

Compartir

Este documento será compartido a través de diferentes medios, a modo de portafolio de trabajo del analista, autor de este trabajo. Google Drive, Kaggle y GitHub serán las plataformas donde alojen este documento, el *scrip* con extensión R, así como el archivo R Markdown en sus versiones HTML y PDF.

En cuestión, el archivo R Markdown, "CPADG_Proyecto1_Programación_R", representará el código con el que se resolvieron las preguntas empresariales y también funcionará como la presentación de resultados a los *stakeholders* a través de las visualizaciones de datos y de las conclusiones del trabajo, así como las recomendaciones para el negocio basadas en datos.

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.