UNIVERSIDAD CEU SAN PABLO

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

MINERÍA DE DATOS APLICADO A LOS NEGOCIOS: ESTUDIO DEL ANÁLISIS DE LA CESTA DE LA COMPRA CON EL PAQUETE ARULES

DATA MINING APPLIED TO BUSINESS: STUDY OF MARKET BASKET ANALYSIS WITH ARULES PACKAGE

TRABAJO DE FIN DE GRADO

GRADO EN INTELIGENCIA DE LOS NEGOCIOS

AUTOR: ELENA SANTAMARÍA IZQUIERDO

DIRECTOR: PABLO ARÉS GASTESI

DEPARTAMENTO DE MATEMÁTICAS Y CIENCIAS DE DATOS

Madrid, 10 de enero 2024

ÍNDICE

Resumen

En este trabajo se analiza la cesta de la compra para identificar modelos de comportamiento de los clientes. Se ha partido de una base de datos recopilada de internet para examinar las compras en una tienda física. Usando el algoritmo Apriori, se ha llevado a cabo un análisis estadístico y visual en R, revelando reglas de asociación entre productos. El análisis se ha hecho de manera integral y se ha segmentado para diferentes categorías de consumidores y productos.

PALABRAS CLAVE: Apriori, arules, reglas de asociación, análisis de la cesta de la compra, transacciones, soporte, confianza, elevación, ítems frecuentes, categorías de productos

Abstract

In this work, an analysis of the shopping basket is carried out, with the objective of identifying customer behavior models. A database collected from the Internet was used to examine purchases in a physical store. Using the A-Priori algorithm, a statistical and visual analysis has been carried out in R, revealing association rules between products. The analysis was carried out comprehensively and has also been segmented for different categories of consumers and products.

KEYWORDS: Apriori, arules, association rules, market basket analysis, transactions, support, confidence, elevation, frequent items, product categories

1. Introducción

1.1 Cesta de la compra

Márquez Sánchez (2018) apunta que la ciencia y la innovación tecnológica son fundamentales para el crecimiento económico y la productividad. El conocimiento se ha convertido en un recurso incalculable, ocupando un espacio cada vez mayor en todos los ámbitos, especialmente en el económico. En la era actual de globalización, uno de los mayores desafíos es convertir la información en conocimiento útil. Esto implica el desarrollo de las capacidades y habilidades de individuos y organizaciones se convierta en un factor clave para el cambio. Las condiciones actuales del mercado y la búsqueda constante del aumento de ingresos impulsan a las empresas en la búsqueda de alternativas que les permitan crecer y superar los efectos de la competencia global

Para hacer crecer los ingresos de un negocio se requiere que los clientes compren el mayor número de productos posible, la mayor cantidad de un producto específico o ambos. Una manera de conseguir esto es sugerirle más productos. Y la pregunta que se plantea ahora es ¿qué producto sugiero al cliente? Pero para responder a esta pregunta es necesario plantearse antes ¿qué probabilidad hay de que un cliente compre el producto A si ya tiene el producto B en su cesta de la compra? Esto se contesta con el análisis de la cesta de la compra o análisis de asociación, que se realizará en este proyecto.

"*El análisis de la cesta de la compra es una técnica de modelado que busca las relaciones entre los artículos que compran los clientes"* (La, 2018). Es decir, que si un cliente compra un móvil es muy probable que compre una funda y un protector de pantalla, a estos se les denomina productos complementarios[1]. Tener conocimiento sobre esto supone una clara ventaja competitiva sobre otras empresas, ayudando a los comerciantes a averiguar en qué productos es más inteligente aplicar descuentos. Gracias a esta técnica se pueden

[1] Un bien complementario es aquel que se debe utilizar juntamente con otro para poder satisfacer la demanda del consumidor (Nicole Roldán, 2020)

descubrir patrones desconocidos hasta el momento, siendo uno de los métodos más importantes para descubrir las relaciones entre los artículos.

Para entender cómo funciona esta técnica se utilizan nociones de causa y efecto o antecedentes y consecuentes, a cada uno de los productos que aparecen en una transacción se le denomina ítem, siendo las transacciones las compras realizadas por cada cliente. En el ejemplo mencionado anteriormente, el móvil es el antecedente y la funda y el protector los consecuentes: {Móvil} → {Funda, protector de pantalla}

1.2 Justificación del análisis del proyecto

La motivación para elegir el análisis de la cesta de la compra como tema para mi trabajo es por la importancia que está adquiriendo el marketing junto con el análisis de datos, y cómo está afectando a los consumidores.

El análisis del negocio se basa en aprovechar la información que obtiene la empresa para tomar mejores decisiones en los negocios. Siendo además una técnica de extracción de datos que se puede aplicar a muchos ámbitos diferentes, como marketing, bioinformática, educación, etc. El área más afectada ha sido el marketing, de ahí surge el nombre *marketing analytics*, que, respecto al análisis de datos, es el proceso de recolección, análisis y ejecución de los *insights* que fueron deducidos de los grandes bases de datos para acercarse a las necesidades del cliente y mejorar los resultados del marketing y la contabilidad interna de la empresa (Rivera, 2015). Por lo tanto, un mayor uso del *marketing analytics* provoca un aumento de beneficios y ventas

2. Comprensión de los datos

2.1 Recolección de los datos

Los datos utilizados en este trabajo son de “The Complete Journey” y han sido extraídos de la web “Dunnhumby”, una página online que usa datos sobre la experiencia de los consumidores los interpreta y toma decisiones inteligentes en tiempo real, con el fin de conseguir una experiencia más satisfactoria tanto para las marcas como para los consumidores. En enlace para descargar los datos es

https://www.dunnhumby.com/source-files/ donde se pueden encontrar los archivos en formato csv y sas. La información que aporta es de transacciones o compras de 2,500 hogares en dos años antes de 2015 y compradores minoristas frecuentes.

En la Figura A.1 (Apéndice **7.1**), se muestran las ocho bases de datos que contiene el archivo: “campaign_table” (Tabla de la Campaña), “campaign_desc” (Descripción de la Campaña), “coupon_redempt” (Cupones Utilizados), “hh_demographic” (Demografía Familiar), “coupon” (Cupones), “transaction_data” (Datos de Transacciones), “product” (Producto) y “casual_data” (Datos Frecuentes).

Para realizar el análisis de la cesta de la compra las bases de datos empleadas en este estudio son: “hh_demographic”, “transaction_data”, “product” y “coupon”. El resto de los datos acerca de las campañas realizadas no se han tenido en cuenta en el estudio

2.2 Selección de datos y descripción

En "transaction_data” hay un total de 92,339 productos, que se identifican a través del ID y las variables que se van a usar son HOUSEHOLD_KEY, PRODUCT_ID, QUANTITY y COUPON_DISC

La base de datos consta de las siguientes variables:

Tabla 1: Base de datos de las transacciones

HOUSEHOLD_KEY	Identificador del hogar, se usa para unir las tablas
BASKET_ID	Identificador de la compra, necesario para aplicar el algoritmo Apriori en este estudio
DAY	Día
PRODUCT_ID	Identificador del producto, se usa para unir las tablas
QUANTITY	Cantidad
COUPON_DISC	Descuento aplicado por el fabricante del cupón

Por otro lado “hh_demographic” contiene la situación demográfica de 801 hogares diferenciándolos por el “HOUSEHOLD_KEY”. Para el análisis del algoritmo Apriori se usan las variables: “INCOME_DESC” Y “MARITAL_STATUS_CODE”

Tabla 2: Base de datos de la demografía familiar

HOUSEHOLD_KEY	Identificador del hogar
INCOME_DESC	Ingresos
AGE_DESC	Edad
HOMEOWNER_DESC	Tipo de residencia: alquilada, propietarios, de alquiler, propietarios o de propiedad desconocida
MARITAL_STATUS_CODE	Estado matrimonial: casado, soltero, desconocido
HOUSEHOLD_SIZE_DESC	Número de individuos del hogar
HH_COMP_DESC DESC	Composición familiar: mujer soltera, hombre soltero, dos adultos sin hijos, dos adultos con hijos, un adulto con hijos o situación desconocida
KID_CATEGORY_DESC	Número de hijos: desconocido, 1, 2, 3+

En la base "product" se van a usar las variables COMMODITY_DESC y SUB_COMMODITY_DESC, estos junto con el BASKET_ID son necesarios para aplicar el algoritmo Apriori (se aplica la categoría o subcategoría según lo que se estudie).

Tabla 3: Base de datos de producto

PRODUCT_ID	Identificador del producto
COMMODITY_DESC	Grupos de productos similares a menor nivel
SUB_COMMODITY_DESC	Grupos de productos similares al nivel más bajo
CURR_SIZE_OF_PRODUCT	Tamaño del producto

Por último, la base de "coupon", que contiene los detalles de los cupones entregados a los clientes basados en sus preferencias en las anteriores compras (no tienen por qué haberlos usado). En este caso se van a usar las tres variables disponibles.

Tabla 4: Base de datos de los cupones

PRODUCT_ID	Identificador de cada producto
COUPON_UPC	Identificador de cada cupón
CAMPAIGN	Identificador de cada campaña del 1 al 30

3. Metodología

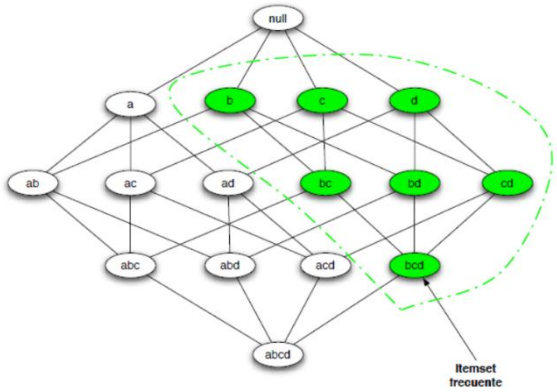
3.1 Herramientas y técnicas utilizadas

La evaluación de este proyecto demanda el empleo de una base de datos extensa para proporcionar un enfoque más integral y preciso. Cuando se manejan grandes volúmenes de datos, como es el caso, es esencial usar herramientas especializadas en estadística, se ha optado por usar el lenguaje de programación R, y su interfaz gráfica RStudio.

Aunque existen varios métodos para realizar un análisis de la cesta de la compra, en este análisis se va a utilizar el algoritmo Apriori, ideado en 1994 por Agrawal y Srikant. Es una técnica de minería de datos que busca la forma más eficiente para obtener "conjuntos de ítemsets[2] frecuentes", que se utilizan posteriormente para crear las reglas de asociación.

Los pasos para seguir del algoritmo Apriori son los siguientes:

1. Obtención de los ítems frecuentes: con un soporte mayor al mínimo establecido por el usuario. Además, se sabe que todo subconjunto de un conjunto de ítems frecuentes también será un conjunto de ítems frecuentes
2. Tras tener seleccionados los ítems frecuentes, el algoritmo genera muchas combinaciones de conjuntos de ítems y busca conjuntos de ítems frecuentes. **Gráfico 1:** Generación de combinaciones de itemsets frecuentes (Pang-Ning Tan, Michael Steinbach, Vipin Kumar 2006)



[2] Un itemset es un conjunto de uno o más ítems (Addison-Wesley, 2006)

Para tener una visión más clara del funcionamiento del algoritmo se va a ejemplificar con la siguiente matriz (donde u son los productos y T son las transacciones)

u1= portátil, u2= móvil, u3= auriculares, u4=cargador, u5=impresora

	u1	u2	u3	u4	u5
T1	1	0	1	0	0
T2	0	1	1	1	0
T3	1	0	1	1	1
T4	0	1	0	1	0
T5	1	0	0	1	1

Transacción 1: {portátil, auriculares}

Transacción 2: {móvil, auriculares, cargador}

Transacción 3: {portátil, auriculares, cargador, impresora}

Transacción 4: {móvil, cargador}

Transacción 5: {portátil, cargador, impresora}

Es aconsejable que se defina un valor bajo para el soporte y un valor elevado para la confianza. Así, se van a crear muchas reglas desde el inicio, ya que se parte con muchos ítems, y con la confianza se seleccionan las más relevantes. *"Una regla de asociación con un valor de confianza bajo no expresará un patrón de comportamiento en los datos y, por otra parte, un valor de soporte muy elevado probablemente llevaría a la perdida de patrones" (Neves, 2008).* Hay que tener en cuenta que, a pesar de ser un algoritmo muy útil y ser muy utilizado, su ejecución es muy costosa, ya que se realizan muchas búsquedas y combinaciones.

3.2 Medidas de asociación

Es de gran importancia entender la fuerza de las reglas de asociación entre los distintos elementos (Fernández Sandoval, 2014). Se pueden clasificar en:

1. Reglas accionables: cuando se dispone de gran cantidad de información y se encuentran varios patrones de asociación. Siendo reglas sencillas con las cuales se toma una decisión clara
2. Reglas triviales: Son patrones muy comunes, que son obvios para cualquier persona que trabaje en el negocio. Estas reglas nos permiten comprobar el correcto funcionamiento del algoritmo, con datos verídicos y resultados válidos, pero no aportan información nueva, son reglas redundantes.

3. Reglas inexplicables: aportan resultados confusos o poco útiles en el estudio, que no aportan una decisión clara

Se van a analizar miles de productos por categorías y subcategorías, por lo que se eligen las medidas de asociación más adecuadas para este análisis de la cesta de la compra, que son: soporte, elevación y confianza.

Soporte: "*Mide la frecuencia con la que se encuentra un itemset en todo el conjunto de transacciones, se utiliza un umbral de soporte específico para reducir la cantidad de itemsets que se van a analizar*" *(Hahsler et al., 2017)*. Al comienzo del análisis se puede usar como umbral de soporte entre el 1% y el 10%. Su fórmula es la siguiente:

$$\text{SOPORTE}(\text{X}) = \frac{\text{FRECUENCIA}(\text{X})}{\text{N}} \text{ (n como número total de transacciones)} \quad \textbf{Ecuación 1}$$

Confianza: "*Dice la probabilidad de que se compre Y dado que se compró X, (X → Y). La confianza se puede interpretar como una estimación de la probabilidad* P(X|Y)*"* (Hahsler et al., 2017). Se calcula de la siguiente forma:

$$Confianza(X \rightarrow Y) = \frac{soporte(X \cup Y)}{soporte\ (X)} \quad \textbf{Ecuación 2}$$

Es necesario considerar que la confianza puede distorsionar la relevancia de una asociación, ya que solo tiene en cuenta la popularidad de X (por ejemplo, móvil) y no la de Y (por ejemplo, cargador). Para abordar la popularidad individual de ambos elementos, emplearemos una tercera medida explicada a continuación.

Elevación o lift: "*Dice cómo de probable es que se compre Y cuando se compra X, es decir, mide cuántas veces más X & Y ocurren juntas si fueran estadísticamente independientes*" (Hahsler et al., 2017). Cuya fórmula es:

$$Elevación(X \rightarrow Y) = \frac{Confianza(X \rightarrow Y)}{Soporte\ (X)} = \frac{soporte(X \cup Y)}{soporte\ (X) \cdot soporte\ (Y)} \quad \textbf{Ecuación 3}$$

La elevación puede adquirir los siguientes valores:

- Lift=1: No hay asociación entre X & Y.
- Lift>1: Es probable que se compre Y si se compra X.
- Lift<1: Es poco probable que se compre Y si se compra X

Continuando con el ejemplo del apartado **3.1**, se calculan las medidas de asociación.

Soporte (u1): $\frac{3}{5}$ = 0.6 (u2): $\frac{2}{5}$ = 0.4 (u3): $\frac{3}{5}$ = 0.6 (u4): $\frac{4}{5}$= 0.8 (u5): $\frac{2}{5}$ = 0.4

Confianza (u1→u2): 0 (u1→u3): $\frac{0.4}{0.6}$ = 0.67 (u1→u4): $\frac{0.4}{0.6}$= 0.67 (u1→u5): $\frac{0.4}{0.6}$= 0.67

Elevación (u1→u2): 0 (u1→u3): $\frac{0.67}{0.6}$ = 1.12 (u1→u4): $\frac{0.67}{0.6}$= 1.12 (u1→u5): $\frac{0.67}{0.6}$=1.12

Nota: solo se ha calculado la confianza y la elevación del primer producto respecto al resto para ver el ejemplo, habría que hacer esto con todos los productos

En el ejemplo dado, el cargador es el producto más repetido en las transacciones, la confianza del portátil con el cargador es mayor al 50% por lo que se cuando un cliente compra un portátil es probable que compre un cargador (también ocurre con la impresora y el móvil). Además, como la elevación es mayor que 1 es probable que ambos productos se compren juntos

4. Preparación de los datos

4.1 Análisis exploratorio y limpieza de datos

En la base de datos de producto hay 30,652 NA de 92,353 filas, lo que supone el 33% de los datos. La mayoría de los NA se encuentran en la variable tamaño del producto dentro de la base de “product”, por lo que, en vez de eliminar todos los registros que con tengan NA en alguna de sus variables, se ha decidido no analizar la variable “CURR-SIZE_OF_PRODUCT” o tamaño del producto.

Al ser encuestas, otro problema que se presenta en estas bases de datos, son las veces que aparece la palabra "None/Unknown", ya que no aporta información relevante al análisis. Las variables en las que aparezca este resultado en cantidades abundantes no se tendrán en cuenta para el estudio, siendo esta el número de hijos con el 70% de respuestas no contestadas

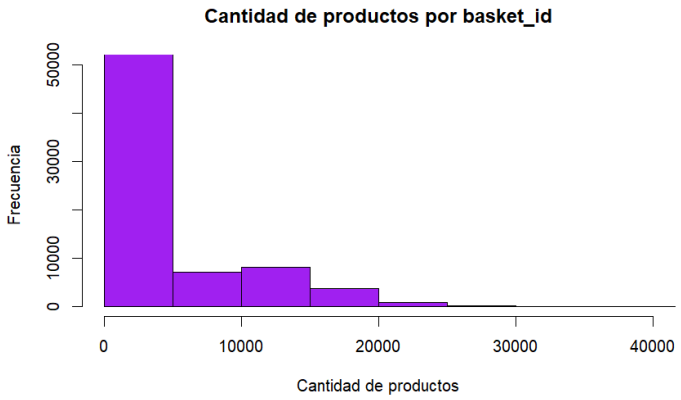
4.2 Análisis descriptivo y estadístico

Tras eliminas los valores nulos de la base de la base de producto (30,652), se procede a hacer un análisis exploratorio de los datos para poder entenderlos, por ello se realizan histogramas (para variables cuantitativas) y gráficos de barras (para variables cualitativas).

De la base de datos de "transaction_data" se representan los histogramas de frecuencias de: cantidad de productos y los productos más comprados.

Sabiendo que las transacciones se repiten varias veces a lo largo de la base de datos, se aprecia una distribución sesgada a la derecha en el gráfico de la cantidad de productos comprados, es decir, la mayor parte de las cantidades se encuentran a la izquierda de la media, lo que indica que la mayoría de los productos se compran en poca cantidad. Esto se confirma al comparar el valor de la media (100.42) con la moda(1), al no ser valores similares la gráfica es asimétrica. También se puede apreciar que las cantidades se mueven en un intervalo entre 1 y 25,000 por cada transacción, no siendo frecuente superar una cantidad mayor a 5,000 productos por transacción.

Gráfico 2: Histograma de la cantidad de productos comprados por los clientes

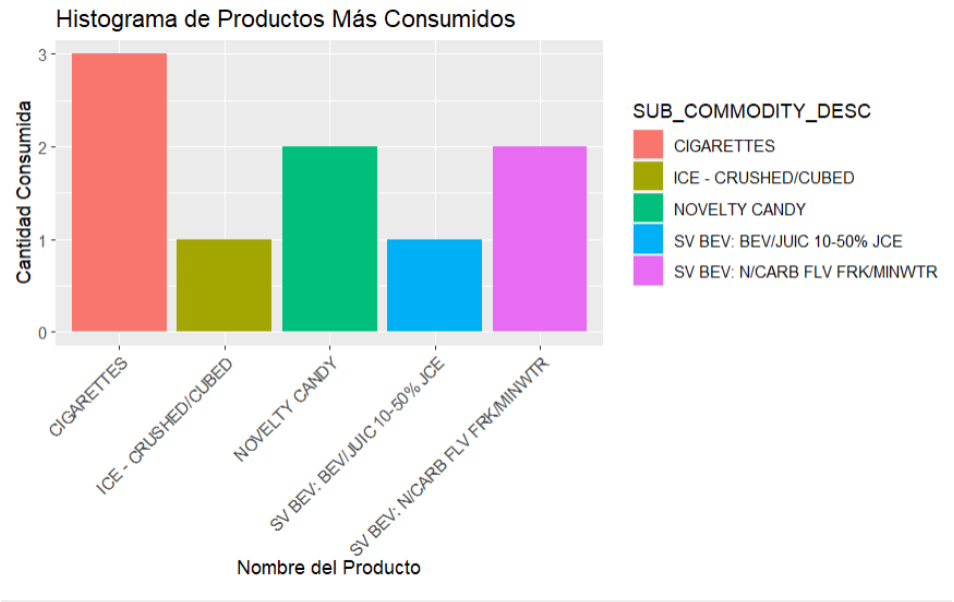


Se ha calculado que los consumidores que mayor cantidad de productos compran varían entre 39,000 y 89,000 productos (en un periodo de dos años), siendo el más consumido la gasolina, que se repite más 30,000 con diferentes PRODUCT_ID en varias transacciones.

Omitiendo la gasolina, se ha hecho un estudio que indica el número de compras por cada producto de los 9 clientes de mayor volumen de consumo. Se descubre que los cigarrillos son el segundo producto más comprado (se repite en tres transacciones de nueve), seguido por los caramelos y las bebidas carbonatadas (repetidos dos veces cada uno),

el hielo y el zumo (repetidos una vez cada uno). Lo que indica que los clientes que más compran son fumadores.

Gráfico 3: Gráfico de barras de los productos más comprados de los nueve clientes con mayor volumen de consumo

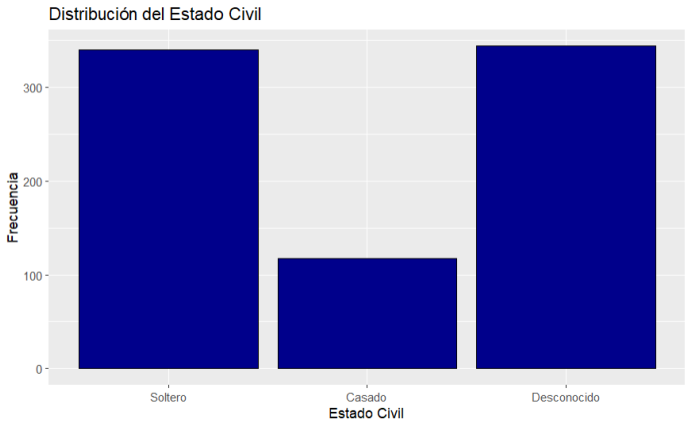


De la base de datos de “hh_demographic” se han utilizado los siguientes diagramas de barras:

En el Gráfico 4 se representa el estado civil de las personas encuestadas, con una cantidad similar entre las personas que no respondieron y las casadas, unas 300 veces, mientras que las solteras representan un tercio de esta cantidad con 117 casos). Esta variable no sería óptima para añadir en el análisis ya que siendo el 43% de los datos desconocidos no sería un estudio fiable[3]. A pesar de ello, el análisis realizado a través del algoritmo en el apartado **5.4**, arroja unos resultados muy valiosos.

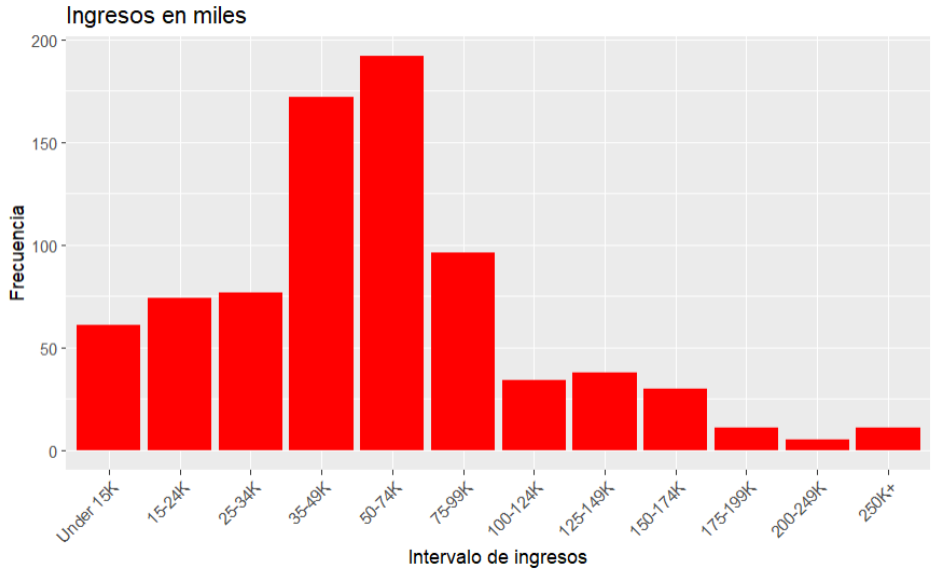
[3] Los datos faltantes aleatorios pueden perturbar el análisis, dado que disminuyen el tamaño de las muestras y, en consecuencia, la potencia de las pruebas de contraste de hipótesis. Los datos faltantes no aleatorios ocasionan, además, disminución de la representatividad de la muestra.

Gráfico 4: Gráfico de barras del estado civil de los clientes



En el caso de los ingresos hay una clara variación entre los rangos, siendo los más repetidos entre 50-74 mil euros y los menos frecuentes los mayores de 200 mil euros, mostrando una figura asimétrica positiva. La media está en 64.24 miles de euros anuales y el coeficiente de variación en 74.05. Con este último dato se comprueba la variabilidad de los datos observada en la gráfica 5

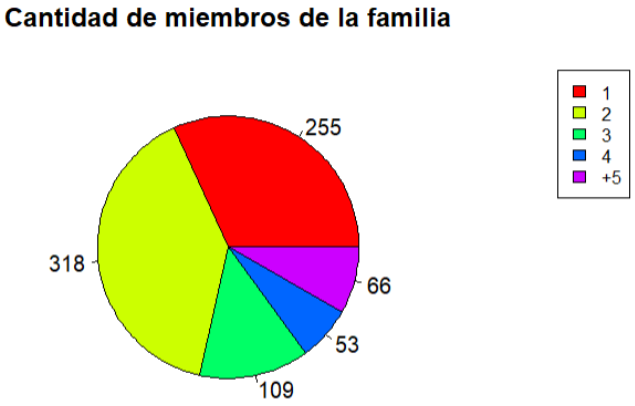
Gráfico 5: Histograma de la distribución de los ingresos



En relación con los miembros que conforman la familia, las parejas conforman son la opción abundante con 318 casos, seguido por los solteros con 255, siendo las familias con hijos las menos frecuentes. Esto se debe tener en cuenta al hacer el análisis ya que

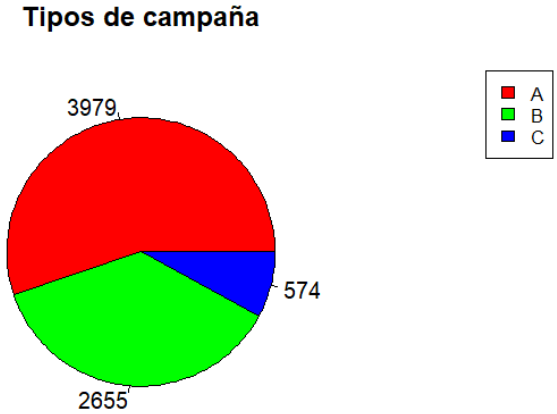
puede suponer que los alimentos para niños sean los menos consumidos y también indica que es más interesante analizar los grupos familiares más frecuentes (solteros y parejas)

Gráfico 6: Gráfico sectorial de los miembros que componen la familia



En el gráfico 7 de la base “campaign_table” se muestran los tres tipos de campaña que se llevan a cabo, A, B y C. No se dispone de una descripción sobre estas, pero se aprecia cómo la campaña A lidera frente a las otras siendo utilizada más del 50% de las veces, por lo que los cupones de esta son los más entregados a los clientes.

Gráfico 7: Gráfico sectorial de los tipos de campaña



En el caso de las variables “COMMODITY” y “SUB_COMMMODITY”, al contener muchos grupos no se han representado gráficamente, pero se han realizado tablas de frecuencias que se pueden consultar en: Tablas de frecuencias

Aunque el resto de las variables no son objeto de estudio de este trabajo, se ha hecho un breve análisis de estas variables, que pueden ser valoradas para utilizadas en un futuro estudio:

- El rango de edad de más consumo es de los adultos entre 45 y 54 años (250 veces), seguido por el rango entre 35 y 44 (200 veces) y en tercera posición entre los 25 y 34 (150 veces), siendo los jóvenes (entre 19 y 24) el grupo minoritario repitiéndose 50 veces.
- Al igual que la variable de estatus matrimonial, la cantidad de hijos que tiene cada familia no es la variable óptima para analizar, ya que el 70% de las personas o no tienen descendientes o prefieren no contestar, y esto afectaría a los resultados obtenidos del estudio
- Los tipos de propietarios de las casas de las personas encuestadas, aunque hay alrededor de 250 personas que no contestaron, 500 de ellas son propietarios de su casa y este es el número más elevado
- La cantidad de miembros que conforma la familia muestra los mismos datos que el gráfico 5 pero más detalladamente, separándolo por sexo y diferenciando si las parejas tienen hijos o no

5. Aplicación del algoritmo Apriori

5.1 Análisis variable "commodity"

Para este apartado se usan 2 bases de datos "transaction_data" y "product", de las cuales se utilizan las columnas "BASKET_ID" y "COMMODITY_DESC".

Como se ha mencionado en el apartado **3.1**, lo primero es saber la frecuencia con la que se repiten los ítems. Esto se hace, primero, a través de la integración de dos o más bases de datos en una tabla con la función *merge*, uniéndolas a través de una clave, en este caso "PRODUCT_ID". Posteriormente se adaptan los datos para que la tabla sólo tenga dos columnas, una que identifique cada transacción o compra y otra para describir las categorías (o subcategorías de los productos, según la variable que se quiera estudiar).

Y por último se aplica la función *itemFrequency* para obtener la frecuencia absoluta con la que se repiten estas categorías o subcategorías de productos

Código 1:

```
tabla_merge2<-merge(transaction_data,product, by="PRODUCT_ID")
datos_split2 <- split(x = tabla_merge2$COMMODITY_DESC, f = tabla_merge2$BASKET_ID)
t2 <- as(datos_split2, Class = "transactions")
frecuencia_items2 <- itemFrequency(x = t2, type = "absolute")
frecuencia_items2 %>% sort(decreasing = TRUE) %>% head(5)
```

Tabla 1: Items más frecuentes de la variable “COMMODITY”

SOFT DRINKS	FLUID MILK PRODUCTS	BAKED BREAD/BUNS/ROLLS	CHEESE	BAG SNACKS
71,699	69,278	60,311	46,898	42,037

Se puede apreciar como los grupos de productos más repetidos son las bebidas no alcohólicas, los lácteos y los productos de pastelería

Para leer los datos de transacciones se utiliza la instrucción *read.transactions* como aparece en el Código 2. En esta función hay que indicar: la ruta del archivo que contiene la nueva tabla con las bases de datos que se van a utilizar, el formato que puede ser *basket* o *single*, (en este caso se ha usado el formato individual ya que las transacciones están desagrupadas), y *cols* que indica las dos columnas que se van a utilizar, una de ellas siempre va a ser “BASKET_ID”, ya que contiene los identificadores de las transacciones.

Código 2:

```
tabla_prod_trans<-merge(transaction_data,product, by="PRODUCT_ID")
write.csv(na.omit(tabla_prod_trans[,c(3,16)]),"C://Users//elesa//Documents//TFG//tabla_merge2.csv", row.names
= FALSE)
transactions4 <- read.transactions("C://Users//elesa//Documents//TFG//tabla_merge2.csv", format = "single", sep
= ",", cols = c("BASKET_ID", "COMMODITY_DESC"), header =TRUE)
inspect(transactions4)
```

Salida del código 2:

```
      items                                transactionID
[1]   {ONIONS,
       ORGANICS FRUIT & VEGETABLES,
       POTATOES,
       TROPICAL FRUIT,
       VEGETABLES - ALL OTHERS}             26984851472
[2]   {BAKED BREAD/BUNS/ROLLS,
       BROOMS AND MOPS,
       COOKIES/CONES,
       PNT BTR/JELLY/JAMS}                  26984851516
[3]   {BEEF,
       BREAKFAST SAUSAGE/SANDWICHES,
       CONVENIENT BRKFST/WHLSM SNACKS,
       CRACKERS/MISC BKD FD,
       EGGS}                                26984896261
[4]   {BAKED BREAD/BUNS/ROLLS,
       SOUP}                                26984905972
[5]   {CANDY - CHECKLANE,
       CANDY - PACKAGED,
       ELECTRICAL SUPPPLIES}                26984945254
```

Como se observa en el Código 3, se define el soporte mínimo, en este caso se ha definido como 30 dividido por la dimensión de las transacciones[4], que da un resultado muy cercano a 0 (0.0001086264), y la confianza se ha establecido en un 0.98. Antes de aplicar el algoritmo, se tienen 276,176 transacciones, y tras aplicarlo quedan 52 reglas. Se pueden ver en la salida del código 3 que, al ordenar las reglas de mayor a menor confianza, las seis primeras tienen una confianza de 1, aunque el soporte de todas ellas es muy bajo. La elevación mínima del todas las transacciones es de 3.9 y la máxima de 47 (siendo mayor a 1 lo que indica que la asociación no se da por causalidad), por tanto, las asociaciones son bastante reales. Todas ellas tienen al menos 3 antecedentes y 1 solo consecuente.

Código 3:

```
soporte <- 30 / dim(transactions4)[1]
rules4 <- apriori (transactions4, parameter = list(supp = soporte, conf = 0.98))
rules_conf2 <- sort (rules4, by="confidence", decreasing=TRUE)[1:10]
inspect(head(rules_conf2))
```

Salida del código 3:

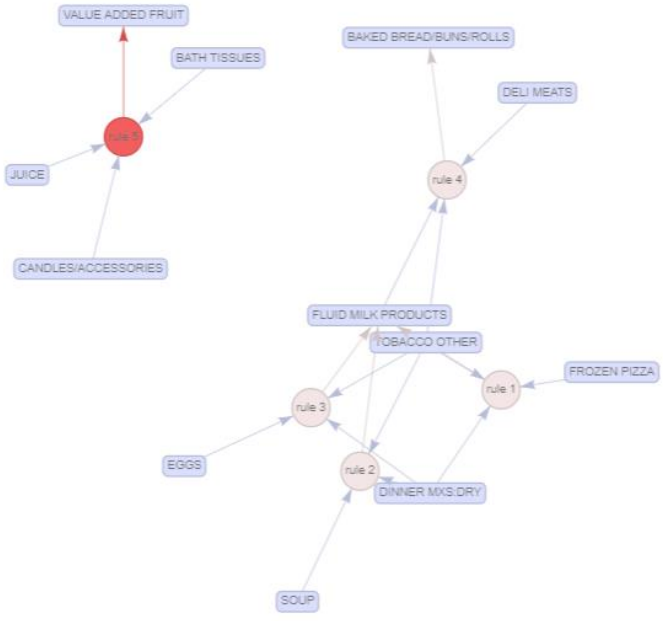
	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	lift <dbl>
[1]	{DINNER MXS:DRY, FROZEN PIZZA, TOBACCO OTHER}	=>	{FLUID MILK PRODUCTS}	0.0001303517	1	3.986489
[2]	{DINNER MXS:DRY, SOUP, TOBACCO OTHER}	=>	{FLUID MILK PRODUCTS}	0.0001194890	1	3.986489
[3]	{DINNER MXS:DRY, EGGS, TOBACCO OTHER}	=>	{FLUID MILK PRODUCTS}	0.0001122473	1	3.986489
[4]	{DELI MEATS, FLUID MILK PRODUCTS, TOBACCO OTHER}	=>	{BAKED BREAD/BUNS/ROLLS}	0.0001629396	1	4.579198
[5]	{BATH TISSUES, CANDLES/ACCESSORIES, JUICE}	=>	{VALUE ADDED FRUIT}	0.0001194890	1	47.145101
[6]	{CANDLES/ACCESSORIES, JUICE, VALUE ADDED FRUIT}	=>	{TROPICAL FRUIT}	0.0001412143	1	8.496677

[4] se está utilizando un enfoque relativo basado en la dimensión total de las transacciones, se ajusta el umbral del soporte en función del tamaño total del conjunto de datos

A continuación, visualizamos las cinco reglas de mayor confianza, para ello se usa la función *plot* del paquete *arulesViz40*.

Cada círculo rojo corresponde con una regla de asociación. Los círculos con un rojo más intenso y un número menor son las más probables que ocurran. Las flechas que apuntan a los círculos son los antecedentes y las que salen de estos los consecuentes. En este caso el círculo más llamativo es rule 5 con una confianza de 1, soporte de 0.000119 y elevación de 47.1, con antecedentes son toallitas de baño, zumo, velas/accesorios y el consecuente es fruta. Esto quiere decir que los clientes compran 47 veces más fruta cuando compran toallitas, zumo y velas. Usando la misma lógica, al tener la primera regla una elevación de 4 es más probable que se compren productos lácteos si ha comprado pizza congelada, tabaco y algún producto relacionado con la cena (no se ha podido identificar el significado de dinner mxs dry). Ambas se pueden considerar reglas inexplicables porque no se entiende la conexión que hay entre los antecedentes y el consecuente, por ello se han modificado el soporte y la confianza mínimos y se ha vuelto a realizar el estudio, con la intención de encontrar reglas más coherentes.

Figura 1: Gráfico de las cinco primeras reglas de la variable “COMMODITY” ordenadas de mayor a menor confianza



Un soporte mínimo apto para realizar este tipo de análisis es de un 1%, al elegir un soporte mayor al anterior se ha reducido la confianza a un 84% (Se puede ver el código

en el apéndice **7.2** código B.1). La mayoría de las reglas con más confianza son triviales y en todas se incluye la categoría de productos lácteos. La conclusión más lógica que se obtiene es que, si se compran productos de bollería o cereales, es 3 veces más probable que se compre leche.

Salida del código 4:

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	lift <dbl>
[1]	{BAKED BREAD/BUNS/ROLLS, CHEESE, COLD CEREAL, EGGS}	=>	{FLUID MILK PRODUCTS}	0.01019640	0.8590604	3.424635
[2]	{CHEESE, COLD CEREAL, EGGS}	=>	{FLUID MILK PRODUCTS}	0.01377745	0.8425598	3.358856
[3]	{BAKED BREAD/BUNS/ROLLS, COLD CEREAL, EGGS}	=>	{FLUID MILK PRODUCTS}	0.01466818	0.8423789	3.358134
[4]	{BAKED BREAD/BUNS/ROLLS, BEEF, CHEESE, COLD CEREAL}	=>	{FLUID MILK PRODUCTS}	0.01155423	0.8266839	3.295567
[5]	{BAKED BREAD/BUNS/ROLLS, CHEESE, EGGS, TROPICAL FRUIT}	=>	{FLUID MILK PRODUCTS}	0.01028330	0.8258215	3.292128
[6]	{CHEESE, FLUID MILK PRODUCTS, HOT DOGS}	=>	{BAKED BREAD/BUNS/ROLLS}	0.01056573	0.8254597	3.779943
[7]	{BAKED BREAD/BUNS/ROLLS, CHEESE, COLD CEREAL, TROPICAL FRUIT}	=>	{FLUID MILK PRODUCTS}	0.01035934	0.8233094	3.282114
[8]	{BAKED BREAD/BUNS/ROLLS, COLD CEREAL, REFRGRATD JUICES/DRNKS}	=>	{FLUID MILK PRODUCTS}	0.01244134	0.8224031	3.278501

8 rows | 1-5 of 8 columns

5.2 Análisis variable subcommodity

El mismo análisis realizado en las categorías se usan en las subcategorías. Con las mismas bases de datos, pero en este caso se seleccionan las columnas “BASKET_ID” y “SUB_COMMMODITY_DESC”. Lo primero es saber la frecuencia con la que se repiten los ítems. La subcategoría más repetida es la leche (61,383), le sigue los plátanos con 30,326 veces y el pan blanco con 26,846.

Código 5:

```
tabla_merge<-merge(transaction_data,product, by="PRODUCT_ID")
datos_split <- split(x = tabla_merge$SUB_COMMODITY_DESC, f = tabla_merge$BASKET_ID)
t <- as(datos_split, Class = "transactions")
frecuencia_items <- itemFrequency(x = t, type = "absolute")
frecuencia_items %>% sort(decreasing = TRUE) %>% head(5)
```

Tabla 2: ítems más frecuentes al analizar la variable “SUBCOMMODITY”

FLUID MILK WHITE ONLY	BANANAS	MAINSTREAM WHITE BREAD	SOFT DRINKS 12/18&15PK CAN CAR	GASOLINE-REG UNLEADED
61,383	30,326	26,846	25,667	24,952

Tras leer las transacciones se establece un soporte mínimo = 50/ número de transacciones (0.000181044) y una confianza del 99%.

Antes de aplicar el algoritmo, se tienen 276,176 transacciones, y tras aplicarlo quedan 53 reglas. Todas las reglas tienen una confianza de 1, aunque el soporte de todas ellas es

cercano a 0. La elevación mínima del total de transacciones es de 4.4 y la máxima de 1.011 (siendo mayor a 1 lo que indica que la asociación no se da por causalidad), por tanto, las asociaciones son muy fiables. A diferencia con la variable “COMMODITY”, aquí se pueden encontrar entre 2 y 4 antecedentes y 1 consecuente

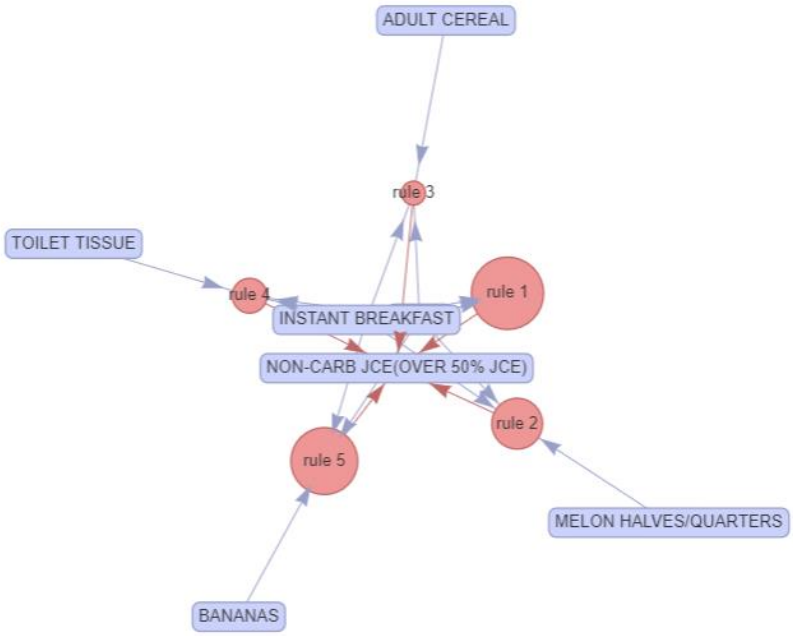
Salida de código 5:

```
     lhs                                  rhs                                 support confidence     coverage        lift count
[1]  {CIDER,
      INSTANT BREAKFAST}                => {NON-CARB JCE(OVER 50% JCE)} 0.0002244945          1 0.0002244945 1011.633700    62
[2]  {CIDER,
      INSTANT BREAKFAST,
      MELON HALVES/QUARTERS}            => {NON-CARB JCE(OVER 50% JCE)} 0.0002100110          1 0.0002100110 1011.633700    58
[3]  {ADULT CEREAL,
      CIDER,
      INSTANT BREAKFAST}                => {NON-CARB JCE(OVER 50% JCE)} 0.0001919066          1 0.0001919066 1011.633700    53
[4]  {CIDER,
      INSTANT BREAKFAST,
      TOILET TISSUE}                    => {NON-CARB JCE(OVER 50% JCE)} 0.0001991484          1 0.0001991484 1011.633700    55
[5]  {BANANAS,
```

En este caso todos los círculos tienen la misma intensidad ya que la elevación es la misma, de 1.001. Al ser la elevación igual a 1, los productos son independientes.

El producto de la leche (los lácteos en general) no aparece entre las cinco primeras reglas cuando se aplica el algoritmo en las variables “COMMODITY” y “SUBCOMMODITY”, pero mirando el total de las reglas es el producto consecuente más repetido, por lo que puede no ser una categoría muy fiable y en el apartado **5.3** se hará el mismo estudio omitiendo**¡Error! La autoreferencia al marcador no es válida.**

Figura 2: Gráfico de las cinco primeras reglas ordenadas de mayor a menor confianza



5.3 Análisis excluyendo el producto de la leche

Se vuelve a aplicar el algoritmo Apriori, pero eliminando antes de la base de datos los ítems de la leche para ver si varía la confianza o el soporte de las reglas obtenidas anteriormente. Para eliminar la leche no se pueden eliminar las observaciones de este producto, ya que, al eliminar una compra, podría afectar los resultados obtenidos de analizarla y no sería un análisis fiel a la realidad. Por ejemplo, si una regla determina {leche, café} → azúcar, puede que, al eliminar la leche, el cliente ya no compre azúcar si ha comprado café, por lo tanto, esta regla sería falsa y habría que eliminarla. Como el mismo "BASKET_ID" se puede se repite en varias observaciones, hay que eliminar todos los "BASKET_ID" que contengan leche.

Código 6:

```
resultado_subset2 <- subset(tabla_merge, grepl("FLUID MILK WHITE ONLY",
tabla_merge2$SUB_COMMODITY_DESC))
resultado_subset2
# Definir el tipo de producto que deseas eliminar
tipo_a_eliminar2 <- "FLUID MILK WHITE ONLY"
# Utilizar dplyr para filtrar y eliminar los registros
datos_filtrados_subcom <- subset(tabla_merge, !(BASKET_ID %in%
unique(BASKET_ID[tabla_merge2$SUB_COMMODITY_DESC == tipo_a_eliminar2])))
print(datos_filtrados_subcom)
```

Tras eliminar la leche, usamos la nueva tabla con los datos filtrados ("datos_filtrados_subcom") para aplicar el algoritmo, el código aparece en el apéndice **7.2** código B.2. Se establecen los mismos parámetros para el soporte y la confianza para poder comparar la cantidad de reglas obtenidas, pero haciéndolo así no se creaba ninguna regla, por lo que se establece como soporte: 30/dimensión del número de transacciones y confianza: 0.89.

Salida código 7:

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>
[1]	{GRND/PATTY - FROZEN, HOT DOG BUNS, PREPARED BEANS - BAKED W/PORK}	=>	{HAMBURGER BUNS}	0.0002607161	0.9491525
[2]	{HEAD LETTUCE, LEAN, MEXICAN SEASONING MIXES, SOUR CREAMS}	=>	{SHREDDED CHEESE}	0.0002514048	0.9473684
[3]	{GRND/PATTY - FROZEN, HOT DOG BUNS, IWS SINGLE CHEESE}	=>	{HAMBURGER BUNS}	0.0003771073	0.9418605
[4]	{BUTTER SPRAY CRACKER (RITZ/CLU, MISCELLANEOUS, ONIONS OTHER}	=>	{CREAM CHEESE}	0.0003119282	0.9054054
[5]	{GRND/PATTY - FROZEN, HOT DOG BUNS, SMOKED/COOKED}	=>	{HAMBURGER BUNS}	0.0003817629	0.9010989
[6]	{CONDIMENTS/SUPPLIES}	=>	{SALAD BAR FRESH FRUIT}	0.0016155089	0.8989637
[7]	{GRND/PATTY - FROZEN, HOT DOG BUNS, POTATO CHIPS}	=>	{HAMBURGER BUNS}	0.0004841871	0.8965517
[8]	{GRND/PATTY - FROZEN, HOT DOG BUNS, PREMIUM - MEAT}	=>	{HAMBURGER BUNS}	0.0003212395	0.8961039
[9]	{CAKE DECORS - BIRTHDAY CANDLES, LAYER CAKE MIX}	=>	{FROSTING}	0.0002653718	0.8906250
[10]	{ECONOMY - MEAT, GRND/PATTY - FROZEN, HOT DOG BUNS}	=>	{HAMBURGER BUNS}	0.0002653718	0.8906250

Se descubren 10 reglas, todas ellas con una elevación muy alta, por lo que son resultados fiables. Fijándose en la elevación de estas reglas, se pueden sacar las conclusiones de:

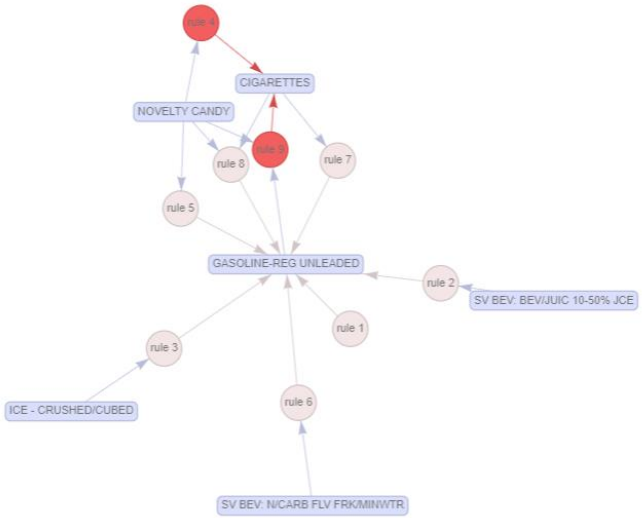
- Si se compra lechuga, salsa mexicana picante y crema agria es 20 veces más probable que se compre queso rayado
- Si se compra carne de hamburguesa congelada, pan de perrito caliente y patatas de bolsa es 34 veces más probable que se compre pan de hamburguesa
- Si se compra decoración de tarta de cumpleaños y un pastel de varias capas es 110 veces más probable que se compre glaseado

Para los próximos análisis con otras variables no se tendrá en cuenta la leche

5.4 Reglas de asociación en las diferentes categorías

Teniendo en cuenta la cantidad de los productos más comprados por los consumidores, explicado en el apartado **2.2**, se ha aplicado el algoritmo a los 13 productos más consumidos (se puede encontrar la tabla en el apartado **7.1** del Apéndice), con una confianza del 0.8 y un soporte mínimo de 0.01. La conclusión a la que se llega es que, si el cliente compra cigarrillos, caramelos o/y helado es un 100% de probable que heche gasolina a su vehículo, (con lo cual se asume que esto ocurre cuando se compra en una gasolinera), siendo igual de probable que tras comprar caramelos se compren cigarrillos. En la figura 3 se pueden ver claramente estas asociaciones y el código de esta se encuentra en apéndice **7.2,** código B.3

Figura 3: representación de las reglas con los productos más comprados



Posteriormente se analizan los datos de los ingresos, agrupados en doce intervalos desde menores de 15 mil euros hasta mayores de 250 mil euros. (Tabla 3). Para tener una visión más clara sobre esta variable, se procede a dividirla en tres categorías: bajo, medio y alto. Según Eurostat (2012), la distribución del ingreso mide la desigualdad de este, y se calcula usando el ingreso total recibido, considerando el 20% de la población con ingresos más altos (el quintil superior) y el 20 % de la población con ingresos más bajos (el quintil inferior). Teniendo esto en cuenta, se calcula que el 20% del total corresponde con 160 individuos, con lo que usando la frecuencia como referencia, se dividen los intervalos de los ingresos quedando de la siguiente manera:

Código 8:

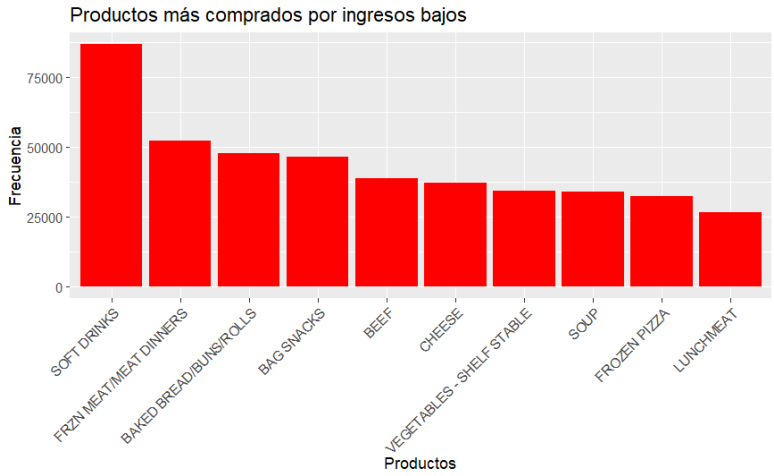
```
tabla_frecuencias3 <- table(demografico$INCOME_DESC)
df_tabla3 <- as.data.frame(tabla_frecuencias3)
df_tabla3$cat<-c("Bajo", "Bajo", "Medio", "Medio", "Medio", "Medio", "Alto", "Alto", "Alto", "Alto", "Alto", "Alto")
colnames(df_tabla3) <- c("Ingresos", "Frecuencia", "Categoría")
kable(df_tabla3, caption = "Tabla de Frecuencias")
df_tabla3
```

Tabla 3: Intervalos de ingresos diferenciados por categorías

Intervalos de Ingresos	Frecuencia	Categoría
< 15K	61	Bajo
15-24K	74	Bajo
25-34K	77	Medio
35-49K	172	Medio
50-74K	192	Medio
75-99K	96	Medio
100-124K	34	Alto
125-149K	38	Alto
150-174K	30	Alto
175-199K	11	Alto
200-249K	5	Alto
250K+	11	Alto

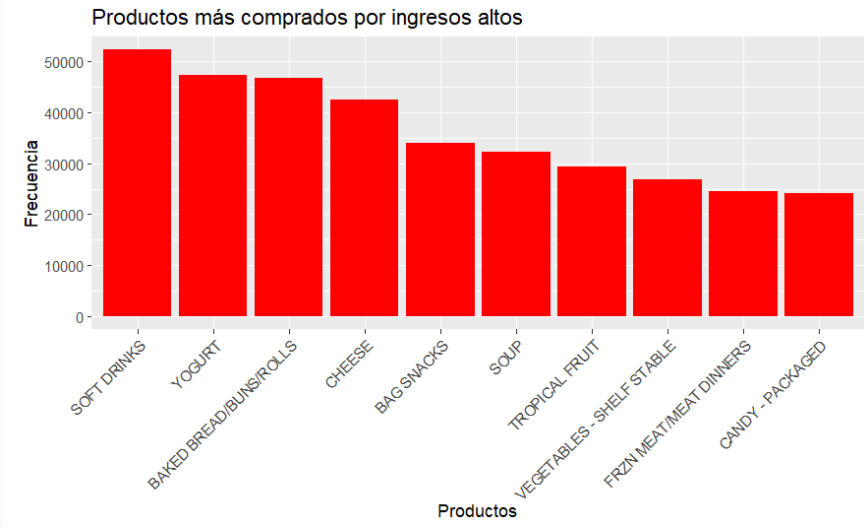
A simple vista se puede ver que los ingresos medios representan el intervalo más grande en concreto entre 35 y 74 mil euros. Se procede a hacer un gráfico de barras con los productos más repetidos, separados por categorías. En la categoría de ingresos bajos destacan las bebidas sin alcohol con más de 75,000 productos, la carne congelada, productos de panadería y snacks (alrededor de 50,000 productos). Estos productos parecen los más comprados, baratos y sencillos de cocinar.

Gráfico 8: Histograma de los ingresos bajos de la variable “COMMODITY”



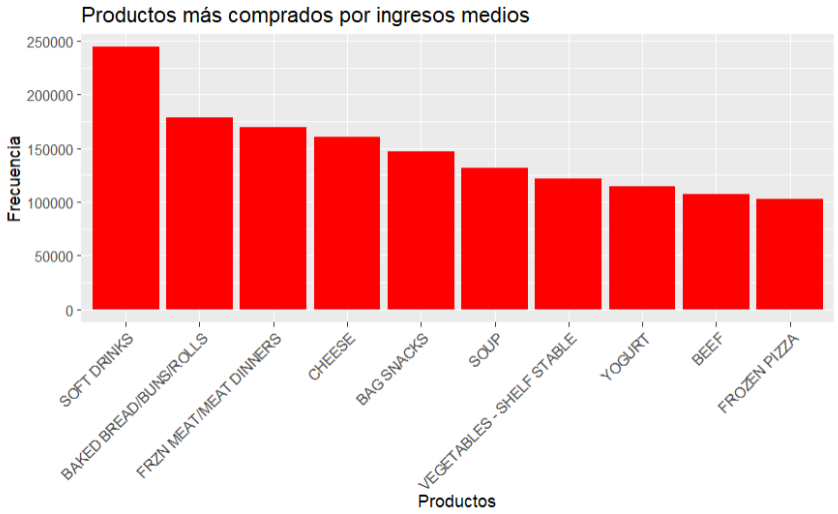
Por el contrario, en los ingresos altos la cantidad de productos comprados no es tan variable, rondan entre 2,000 y 5,000. Aunque la variación entre los ingresos elevados y bajos sea significativa, coinciden varias categorías de productos en ambas gráficas; verduras de larga duración, carne congelada, bebidas sin alcohol, panadería, snacks de bolsa y queso. La principal diferencia es el orden en el que aparecen estas categorías, en este caso, para ingresos altos, el queso aparece en cuarta posición y la carne congelada en penúltima, mientras que en ingresos bajos esta última se sitúa en segunda posición y el queso en mitad de la gráfica. Otro punto que destacar es que en esta gráfica aparece la fruta tropical (sandía, mango, papaya, coco etc) que están considerados entre los alimentos más caros. (Márquez Sánchez, 2020).

Gráfico 9: Histograma de ingresos altos de la variable “COMMODITY”



Por último, se visualizan los ingresos medios, con un gráfico bastante similar al de los ingresos bajos tanto en el orden como en el tipo de producto, aunque las cantidades son mayores que en los dos anteriores gráficos, siendo el producto menos frecuente la pizza congelada, consumida alrededor de 100,000 veces y las bebidas no alcohólicas sobre 250,000

Gráfico 10: Histograma de ingresos medios de la variable “COMMODITY”



Se puede encontrar el código utilizado para calcular la frecuencia de los diferentes intervalos de ingresos en el Apéndice **7.2**, **el** código B.4.

Como el análisis por categorías aporta una información general, se ha decidido aplicar lo mismo para las subcategorías, ya que este contiene información más detallada sobre el tipo de productos. Hay que tener en cuenta que al haber una desagregación de las categorías de los productos la frecuencia disminuye drásticamente.

Como se ha visto en el gráfico anterior, el grupo más frecuente son las bebidas no alcohólicas, y esto se muestra en las dos primeras categorías. También se puede ver que, al tener categorías más desagregadas aparecen productos como los yogures individuales o los plátanos entre las opciones más frecuentes, que en el gráfico de categorías no aparecían, ambos productos no tienen un precio medio elevado, por lo que es coherente con el rango de ingresos estudiado. En los ingresos medios y altos los gráficos son bastante similares, ambos contienen la comida premium congelada y sopa, y en los altos también se dan los productos premium

Gráfico 11: Gráfico de barras de los ingresos altos de la variable “SUBCOMMODITY”

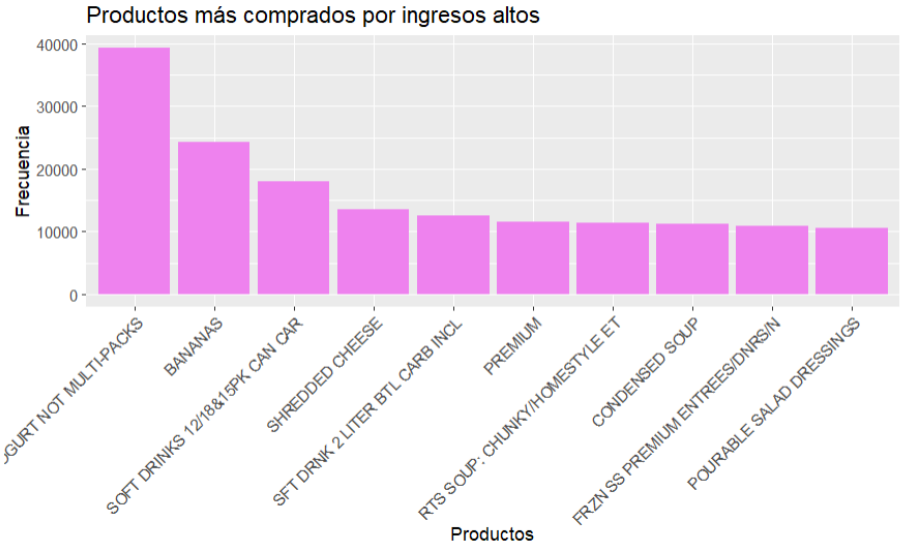
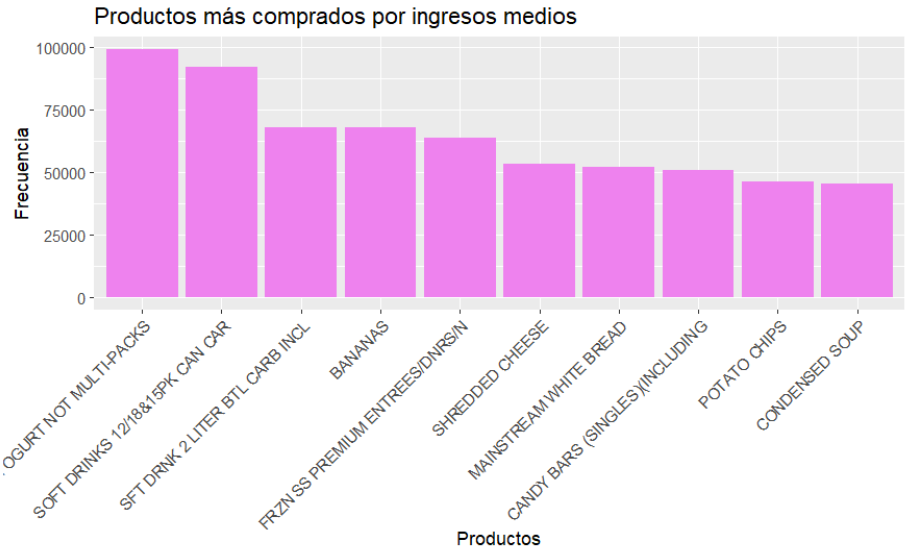


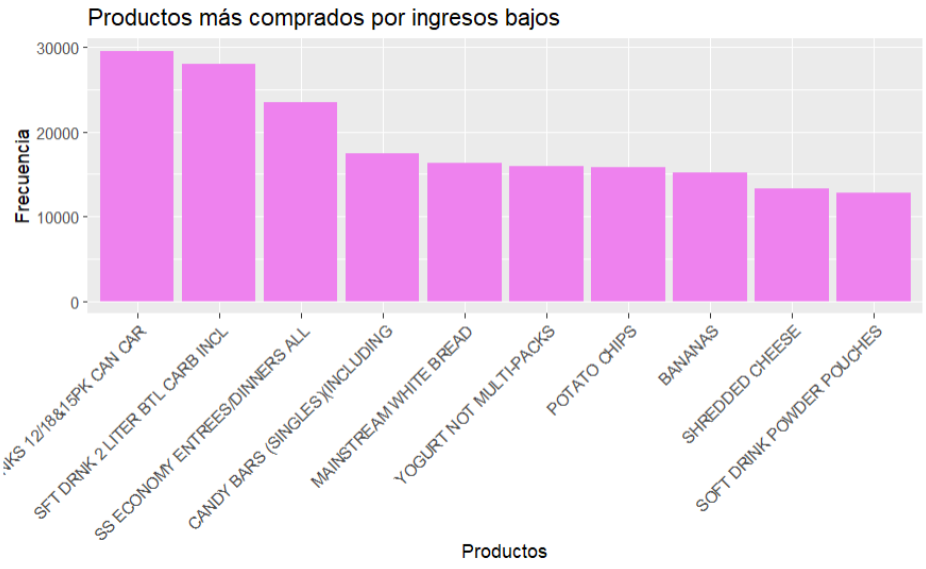
Gráfico 12: Gráfico de barras de ingresos medios de la variable “SUBCOMMODITY”



Al aplicar el algoritmo en los ingresos medios (Gráfico 12) de categoría general, se establece una confianza del 60% y un soporte mínimo de 0.001, con esto se crean siete reglas. La regla con mayor confianza (76%) dice que la compra de fruta tropical y las verduras incita a que los clientes compren productos de panadería, estando presente esta última subcategoría en seis de las siete primeras reglas. En el caso de los ingresos bajos (Gráfico 13) la confianza se ha elevado al 90% , esto significa un margen más

pequeño. De estas se puede concluir que cualquier combinación con verduras, bebidas no alcohólicas, snacks, cerdo, queso, patatas y productos de panadería llevan a los consumidores a comprar carne de vaca en más de un 90% de las veces. Para el análisis de los ingresos altos se ha usado una confianza del 86%, y el resultado final es similar al de los productos de ingreso medio, ya que, mezclando combinaciones de varias categorías (perritos calientes, ensaladas mixtas, queso o sopa entre otros) aparecen los productos de panadería

Gráfico 13: Gráfico de barras de los ingresos bajos de la variable “SUBCOMMODITY”



Otra variable que ha parecido interesante analizar es el estatus matrimonial, comparando las categorías de productos de personas casadas y solteras. (Código disponible en el apéndice **7.2,** código B.5)

Se han utilizado las reglas con mayor confianza (80-95%) y un soporte de 30 entre el número total de transacciones. Comenzando con las personas casadas, el resultado de la mayoría de las reglas es que las parejas acaban aumentando la compra de los productos de panadería en un 96% si antes han comprado; snacks de bolsa, queso, perrito caliente y pizza congelada. La regla con mayor elevación es la 4, cuando los clientes compran snacks de bolsa lo hacen junto con pizza congelada, queso, galletas crackers y salchichas y sándwiches más veces, que sí lo comprasen de forma separada.

Salida del código 9:

```
    lhs                                rhs                          support confidence     coverage      lift c
[1] {BAG SNACKS,
     DELI MEATS,
     HEAT/SERVE}                => {BAKED BREAD/BUNS/ROLLS} 0.0007554201  0.9677419 0.0007806008  7.306452
[2] {BAG SNACKS,
     HOT DOGS,
     SALD DRSNG/SNDWCH SPRD}    => {BAKED BREAD/BUNS/ROLLS} 0.0007806008  0.9687500 0.0008057815  7.314063
[3] {CHEESE,
     HOT DOGS,
     SALAD MIX}                 => {BAKED BREAD/BUNS/ROLLS} 0.0010324075  0.9534884 0.0010827689  7.198837
[4] {BREAKFAST SAUSAGE/SANDWICHES,
     CHEESE,
     CRACKERS/MISC BKD FD,
     FROZEN PIZZA}              => {BAG SNACKS}             0.0007806008  0.9687500 0.0008057815 10.392212
```

Se puede apreciar como las parejas compran comida más pobre nutricionalmente, véase la figura en el apéndice **7.1, Figura 2.**

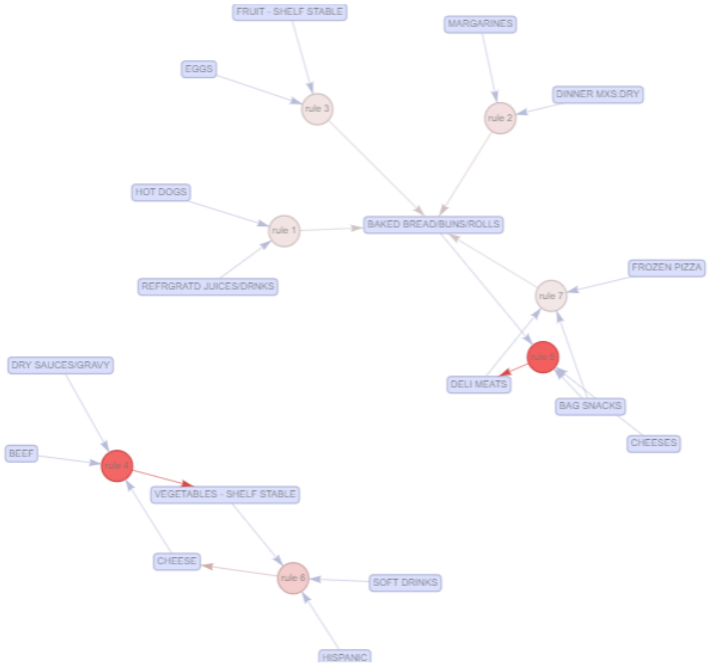
Sin embargo, en las personas solteras hay un ligero cambio. De siete reglas con mayor confianza, la cuarta regla establece que se consume un 82% más de verduras cuando se compra carne de vacuno, queso y salsas, además es la segunda regla con mayor elevación. De este análisis se concluye que las personas solteras tienen una alimentación algo más saludable que las casadas. (Véase el código B.5 **en el apéndice 7.2**)

Salida del código 10:

	lhs <chr>	<chr>	rhs <chr>	support <dbl>	confidence <dbl>
[1]	{HOT DOGS, REFRGRATD JUICES/DRNKS}	=>	{BAKED BREAD/BUNS/ROLLS}	0.002349314	0.8421053
[2]	{DINNER MXS:DRY, MARGARINES}	=>	{BAKED BREAD/BUNS/ROLLS}	0.002349314	0.9142857
[3]	{EGGS, FRUIT - SHELF STABLE}	=>	{BAKED BREAD/BUNS/ROLLS}	0.003083474	0.8400000
[4]	{BEEF, CHEESE, DRY SAUCES/GRAVY}	=>	{VEGETABLES - SHELF STABLE}	0.002496146	0.8292683
[5]	{BAG SNACKS, BAKED BREAD/BUNS/ROLLS, CHEESES}	=>	{DELI MEATS}	0.002863226	0.8297872
[6]	{HISPANIC, SOFT DRINKS, VEGETABLES - SHELF STABLE}	=>	{CHEESE}	0.002349314	0.8421053
[7]	{BAG SNACKS, DELI MEATS, FROZEN PIZZA}	=>	{BAKED BREAD/BUNS/ROLLS}	0.002422730	0.8048780

7 rows | 1-6 of 8 columns

Figura 4: representación se las siete primeras reglas de las personas solteras



Asimismo, se ha decidido aplicar el algoritmo en los cupones de descuento. Para cada tipo de campaña (A, B o C) existen dieciséis cupones específicos, creados en base a las preferencias de los consumidores. Por ello, se han analizado los cupones de cada tipo de campaña, en los tres análisis se fijado una confianza de 0.77 y un soporte de 30 entre el número total de transacciones.

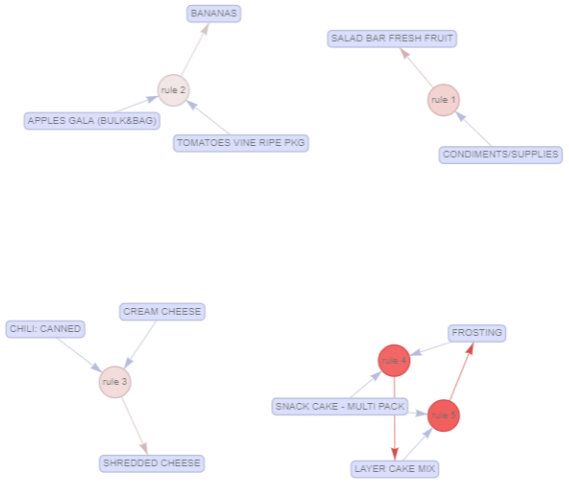
Comenzando por la campaña de tipo A, que es la más abundante y por lo tanto la que más se ajusta a los gustos de la mayoría de los consumidores (utilizada en 3,979 ocasiones, más de un 50% de las veces). Se obtienen cinco reglas, de las cuales las reglas cuatro y cinco son las más parecidas y predecibles, ya que, tanto los antecedentes como los consecuentes incluyen los ingredientes para cocinar una tarta, además una de las reglas es redundante porque aparecen los mismos productos en ambas. Por todo esto la elevación de estas reglas es muy alta, de más de 100 puntos. Incrementando un 83% la compra de pasteles de capas mixtas cuando se compra glaseado y pastelitos. Por otro lado, la regla con mayor confianza es la primera, cuando al comprar condimentos se procede a comprar macedonia de frutas, en un 94% más de probabilidad. (Código B.6 en apéndice **7.2**)

Salida del código 11:

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>
[1]	{CONDIMENTS/SUPPLIES}	=>	{SALAD BAR FRESH FRUIT}	0.0015669290	0.9468085
[2]	{APPLES GALA (BULK&BAG), TOMATOES VINE RIPE PKG}	=>	{BANANAS}	0.0005281783	0.8571429
[3]	{CHILI: CANNED, CREAM CHEESE}	=>	{SHREDDED CHEESE}	0.0005633902	0.7804878
[4]	{FROSTING, SNACK CAKE - MULTI PACK}	=>	{LAYER CAKE MIX}	0.0005457843	0.8378378
[5]	{LAYER CAKE MIX, SNACK CAKE - MULTI PACK}	=>	{FROSTING}	0.0005457843	0.7948718

5 rows | 1-6 of 8 columns

Figura 5: representación de las cinco primeras reglas de la campaña A



Continuando con la campaña B, esta es la segunda más repetida, usada en 2,655 ocasiones. El resultado es muy similar al de la campaña A (hay cuatro reglas iguales), pero en este caso se extraen siete reglas, de las cuales las cuatro últimas tienen una elevación del 100 o cercano a ese valor. Fijándose en los productos, se puede ver que cuando un cliente compra glaseado, es entre un 77 y un 85% probable que compre pasteles de capas mixtas y viceversa. (Código B.7 en apéndice **7.2**)

Salida del código 12:

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>
[1]	{CONDIMENTS/SUPPLIES}	=>	{SALAD BAR FRESH FRUIT}	0.0014933397	0.9493671
[2]	{APPLES GALA (BULK&BAG), TOMATOES VINE RIPE PKG}	=>	{BANANAS}	0.0005973359	0.8571429
[3]	{POTATO CHIPS, TOMATOES VINE RIPE PKG}	=>	{SFT DRNK 2 LITER BTL CARB INCL}	0.0009159150	0.8070175
[4]	{FROSTING, SNACK CAKE - MULTI PACK}	=>	{LAYER CAKE MIX}	0.0005973359	0.8571429
[5]	{LAYER CAKE MIX, SNACK CAKE - MULTI PACK}	=>	{FROSTING}	0.0005973359	0.8571429
[6]	{LAYER CAKE MIX, SHREDDED CHEESE}	=>	{FROSTING}	0.0006968919	0.7777778
[7]	{FROSTING, SOFT DRINKS 12/18&15PK CAN CAR}	=>	{LAYER CAKE MIX}	0.0010353822	0.7761194

7 rows | 1-6 of 8 columns

Por último, se han analizado los resultados de la campaña C, que incluyen productos bastante diferentes a las dos anteriores, únicamente se repite la primera regla. Además, al ser la campaña menos elegida por los consumidores, se ha tenido que reducir la confianza al 0.65 para que se encontraran varias reglas. Entre estas, las de mayor elevación son la cuatro (si se compra acondicionador es 57% más probable que se compre champú), la cinco (si se compra maíz y comida popular es 46% más probable que se compren judías verdes) y la siete (si se compran cereales para niños y espagueti es 41% más probable que se compren alimentos populares[5]). También se informa de que, cuando se compran cereales y barritas multi cereales, es un 12% más probable que se compren plátanos. En general los productos que aparecen en esta campaña son más interesantes nutricionalmente que en las anteriores, por lo que, basándose en los consumidores que han optado por usar cupones de descuento, tan solo el 7.96% de ellos (campaña C ofrecida a 574 personas de 7,208) opta por comprar comida saludable (La figura aparece en el apéndice **7.1**, figura 4 y el código en apéndice **7.2,** código B.8)

[5] Los alimentos considerados más populares mundialmente son: carne de cerdo, maíz, azúcar, patata, arroz, trigo, leche y tomate (Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), 2016)

Salida del código 13:

	lhs		rhs	support	confidence
	<chr>	<chr>	<chr>	<dbl>	<dbl>
[1]	{CONDIMENTS/SUPPLIES}	=>	{SALAD BAR FRESH FRUIT}	0.001308998	0.9459459
[2]	{PIZZA SAUCE}	=>	{SHREDDED CHEESE}	0.001757798	0.7014925
[3]	{SOFT DRINKS CAN NON-CARB (EXCE}	=>	{SOFT DRINKS 12/18&15PK CAN CAR}	0.002917196	0.6666667
[4]	{HAIR CONDITIONERS AND RINSES}	=>	{SHAMPOO}	0.005161194	0.6831683
[5]	{CORN, MAINSTREAM}	=>	{BEANS GREEN: FS/WHL/CUT}	0.001196799	0.6530612
[6]	{ALL FAMILY CEREAL, FRUIT BOWL AND CUPS}	=>	{BANANAS}	0.001346398	0.6923077
[7]	{KIDS CEREAL, SPAGHETTI DRY}	=>	{MAINSTREAM}	0.001271598	0.7083333
[8]	{ALL FAMILY CEREAL, MAINSTREAM WHEAT/MULTIGRAIN BR}	=>	{BANANAS}	0.001757798	0.7580645
[9]	{IWS SINGLE CHEESE, SNACK CAKE - MULTI PACK}	=>	{MAINSTREAM WHITE BREAD}	0.001570798	0.6562500

6. Conclusiones

En el presente estudio se aplica el algoritmo Apriori en gran variedad de productos, haciendo uso de la función *arules*. El objetivo es aplicarlo en el ámbito del marketing para identificar asociaciones entre grupos de categorías o subcategorías. Además de validar asociaciones de productos que pueden considerarse obvias, debido a su reiterada compra conjunta, se busca la relación entre productos cuya asociación no es tan evidente, es decir, las reglas accionables.

Estudiando las categorías y subcategorías sin filtros, se ha observado como los productos lácteos están presentes en todas las compras de confianza elevada y soporte mínimo asignado. Tras eliminar las transacciones que contienen este tipo de producto, aparecen reglas triviales, como: junto con la tarta los clientes optan por comprar glaseado o que la carne de hamburguesa o el perrito caliente se compran con el pan.

Al fijarse en los productos más consumidos por los clientes, aparece la gasolina como el producto más comprado, lo que indica que muchos consumidores tienen coche. Estudiando estas reglas, se intuye que varias compras se realizan en una gasolinera, ya que los productos comprados son snacks o cigarrillos, que suelen encontrarse en este lugar. En este caso la gasolina es el producto consecuente, pero es el motivo principal por el que se va a una gasolinera, por lo que tiene más sentido reducir un poco el precio de la gasolina, para que los clientes se planteen comprar mayor cantidad de productos que se encuentren disponibles en la tienda.

Sin embargo, al filtrar las subcategorías se pueden encontrar asociaciones más significativas. En el caso de los ingresos medios de la categoría general, se ha encontrado una fuerte asociación entre el queso y el pan con el fiambre, y el yogur con

la fruta tropical, con lo que se podría contemplar la posibilidad de aplicar descuentos en el pan o el yogur, para aumentar la compra de productos más caros como es la fruta tropical o el fiambre.

Al comparar las reglas obtenidas según el estatus matrimonial (soltero o casado), usando entre un 80 y 90% de confianza, se concluye que las personas solteras tienen una alimentación algo más saludable que las casadas. Por lo que habrá que usar diferentes tipos de descuentos en función del nicho de mercado al que se quiera dirigir, las parejas serán propensos a comprar más si tienen descuentos en comida rápida, y las personas solteras aumentarán la compran si se reduce el precio de las verduras, la fruta o el queso

El último filtro aplicado es en los cupones de descuento en función del tipo de campaña asignada a los consumidores (A, B o C). A pesar de que las campañas A y B son las más frecuentes, las reglas obtenidas en la campaña C son de mayor utilidad. Usando una confianza del 65%, se aprecia como los consumidores de la campaña C compran productos más saludables. Sería una buena estrategia reducir el precio de productos como el maíz, arroz, patata o carne de cerdo para aumentar el consumo de las verduras, que es un producto más caro. Otra estrategia sería reducir el precio de los cereales y las barritas para que se comprasen más plátanos.

7. Apéndice

7.1 Apéndice A: Figuras y tablas

Figura A.1: Las bases de datos contenidas en *"The Complete Journey*" y sus interconexiones.

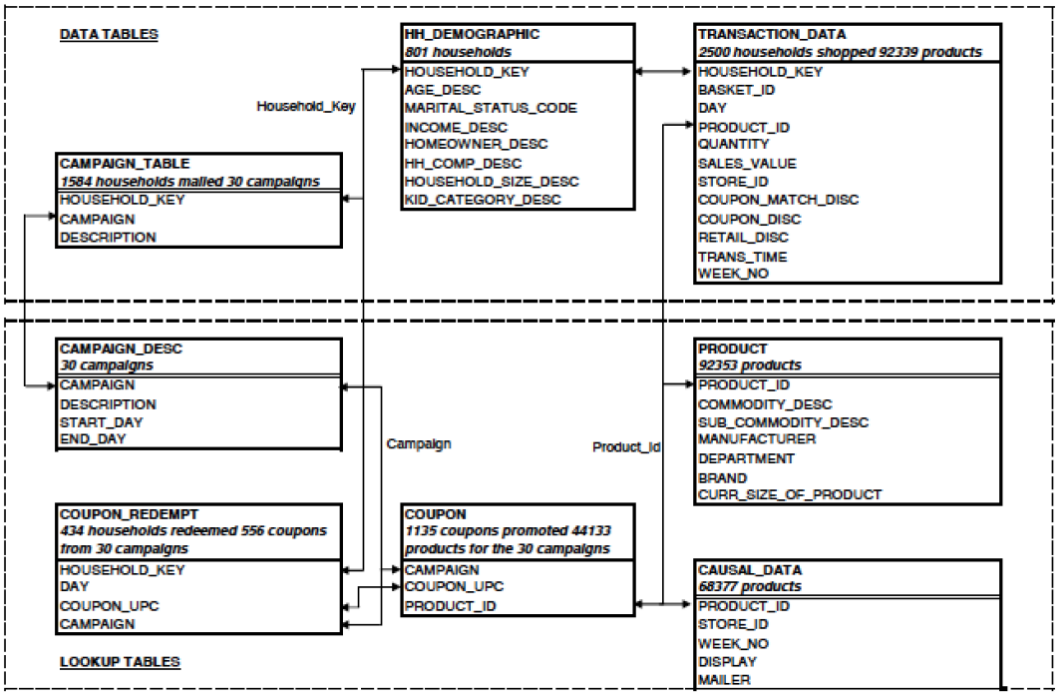


Figura A.2: Representación de las nueve primeras reglas según el estatus matrimonial

Figura A.3: Representación de las nueve primeras reglas de la campaña C

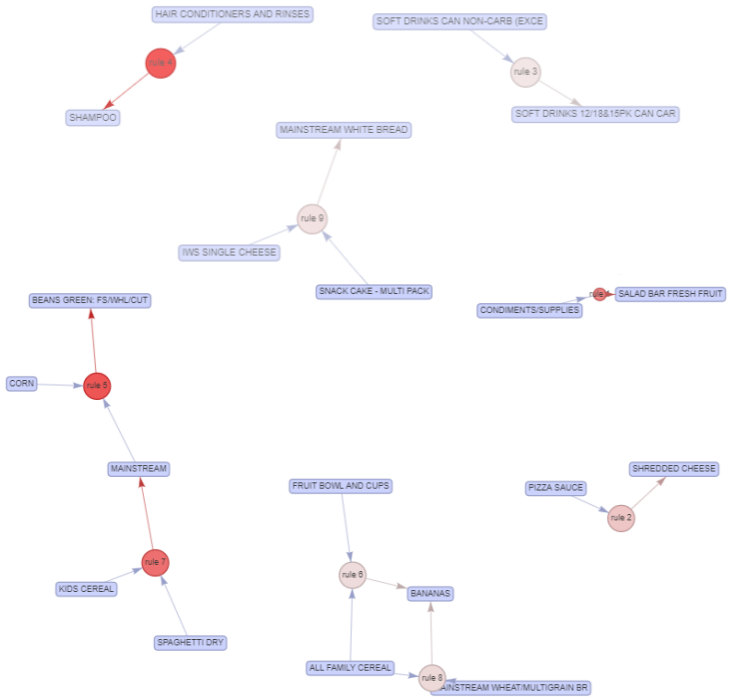


Tabla A.1: Tabla de los diez productos más consumidos

ID DEL PRODUCTO	SUBCATEGORÍA DE LOS PRODUCTOS
1004212	SV BEV: N/CARB FLV FRK/MINWTR
1004212	SV BEV: N/CARB FLV FRK/MINWTR
1049598 C	CANDY
1075368	CIGARETTES
1075368	CIGARETTES
1082990	CIGARETTES
1106186	SV BEV: BEV/JUIC 10-50% JCE
1130777	ICE - CRUSHED/CUBED
5668996	GASOLINE-REG UNLEADED
5716076	GASOLINE-REG UNLEADED

7.2 Apéndice B: Códigos

Código B.1

```
soporte1 <- 0.01
soporte1
rules1 <- apriori (transactions4, parameter = list(supp = soporte1, conf = 0.82))
rules_conf1 <- sort (rules1, by="confidence", decreasing=TRUE)
inspect(rules_conf1)
```

Código B.2

```
write.csv(na.omit(datos_filtrados_subcom[,c(3,17)]),"C://Users//elesa//Documents//TFG//datos_filtra
dos_subcom.csv", row.names = FALSE)
transactions_filtrado2 <-
read.transactions("C://Users//elesa//Documents//TFG//datos_filtrados_subcom.csv", format =
"single", sep = ",", cols = c("BASKET_ID", "SUB_COMMODITY_DESC"), header =TRUE)
#inspect(transactions4)
soporte <- 30 / dim(transactions_filtrado2)[1]
rules_subcom <- apriori (transactions_filtrado2, parameter = list(supp = soporte, conf = 0.89))
```

Código B.3

```
#hacer merge de tabla filtrada con customer transactions
tabla_cantidad<-merge(tabla1_filtrada,top_customers_transactions, by="PRODUCT_ID")
#Creamos csv de la tabla
write.csv(na.omit(tabla_cantidad[,c(8,6)]),"C://Users//elesa//Documents//TFG//tabla_cantidad.csv",
row.names = FALSE)
transactions_cantidad <- read.transactions("C://Users//elesa//Documents//TFG//tabla_cantidad.csv",
format = "single", sep = ",", cols = c("BASKET_ID", "SUB_COMMODITY_DESC"), header =TRUE,
rm.duplicates = TRUE)
cantidad <- 0.01
rules_cant <- apriori (transactions_cantidad, parameter = list(supp = cantidad, conf = 0.8)) # Min
Support as 0.001, confidence as 0.8.
inspect(rules_cant)
#Mostrar la figura de los productos más consumidos por los clientes que más compran
plot(rules_cant, method = "graph", engine="html")
```

Código B.4

```
#Frecuencia de los productos según los ingresos
frecuencia_productos_bajo <- table(tabla_dem_prod_bajo$COMMODITY_DESC)
frecuencia_productos_alto <- table(tabla_dem_prod_alto$COMMODITY_DESC)
frecuencia_productos_medio <- table(tabla_dem_prod2$COMMODITY_DESC)
# Ordenar los productos por frecuencia
productos_ordenados <- sort(frecuencia_productos_bajo, decreasing = TRUE)
productos_ordenados2 <- sort(frecuencia_productos_alto, decreasing = TRUE)
productos_ordenados3 <- sort(frecuencia_productos_medio, decreasing = TRUE)
# Mostrar los 10 productos más comprados de ingresos bajos
prod_ingresos_menores<-head(productos_ordenados, 10)
ggplot(data = data.frame(producto = names(prod_ingresos_menores)),
       aes(x = reorder(producto, -prod_ingresos_menores), y = prod_ingresos_menores)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(x = "Productos", y = "Frecuencia", title = "Productos más comprados por ingresos bajos") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
# Mostrar los 10 productos más comprados de ingresos alto
prod_ingresos_mayores<-head(productos_ordenados2, 10)
ggplot(data = data.frame(producto = names(prod_ingresos_mayores)),
       aes(x = reorder(producto, -prod_ingresos_mayores), y = prod_ingresos_mayores)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(x = "Productos", y = "Frecuencia", title = "Productos más comprados por ingresos altos") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
# Mostrar los 10 productos más comprados de ingresos medios
prod_ingresos_medios<-head(productos_ordenados3, 10)
ggplot(data = data.frame(producto = names(prod_ingresos_medios)),
       aes(x = reorder(producto, -prod_ingresos_medios), y = prod_ingresos_medios)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(x = "Productos", y = "Frecuencia", title = "Productos más comprados por ingresos medios") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Código B.5

```
#aplicamos algoritmo en estatus matrimonial casados
write.csv(na.omit(tabla_dem_prod[,c(1,34)]),"C://Users//elesa//Documents//TFG//tabla_merge3.csv",
row.names = FALSE)
transactions_demografico_producto <-
read.transactions("C://Users//elesa//Documents//TFG//tabla_merge3.csv", format = "single", sep = ",",
cols = c("BASKET_ID", "COMMODITY_DESC"), header =TRUE, rm.duplicates = TRUE)
soporte_demo <- 30 / dim(transactions_demografico_producto)[1]
rules_demo <- apriori (transactions_demografico_producto, parameter = list(supp = soporte_demo,
conf = 0.95))
inspect(rules_demo)
#Mostrar figura con las reglas obtenidas de las personas casadas
plot(rules_demo, method = "graph", engine="html")
#aplicamos algoritmo en estatus matrimonial solteros
write.csv(na.omit(tabla_dem_prod2[,c(1,34)]),"C://Users//elesa//Documents//TFG//tabla_merge4.csv"
, row.names = FALSE)
transactions_demografico_producto2 <-
read.transactions("C://Users//elesa//Documents//TFG//tabla_merge4.csv", format = "single", sep = ",",
cols = c("BASKET_ID", "COMMODITY_DESC"), header =TRUE, rm.duplicates = TRUE)
soporte_demo2 <- 30 / dim(transactions_demografico_producto2)[1]
rules_demo2 <- apriori (transactions_demografico_producto2, parameter = list(supp = soporte_demo2,
conf = 0.8))
inspect(rules_demo2)
#Mostrar figura con las reglas obtenidas de las personas solteras
plot(rules_demo2, method = "graph", engine="html")
```

Código B.6

```
#aplicamos el algoritmo a los cupones A
write.csv(na.omit(tabla_cup_dem[,c(8,22)]),"C://Users//elesa//Documents//TFG//tabla_cupones.csv",
row.names = FALSE)
transactions_cupones <- read.transactions("C://Users//elesa//Documents//TFG//tabla_cupones.csv",
format = "single", sep = ",", cols = c("BASKET_ID", "SUB_COMMODITY_DESC"), header =TRUE,
rm.duplicates = TRUE)
soporte_cup <- 30 / dim(transactions_cupones)[1]
#inspect(transactions3)
rules_cup <- apriori (transactions_cupones, parameter = list(supp = soporte_cup, conf = 0.77))
inspect(rules_cup)
#Mostramos la figura con las reglas obtenidas de la campaña A
plot(rules_cup, method = "graph", engine="html")
```

Código B.7

```
#aplicamos el algoritmo a los cupones B
write.csv(na.omit(tabla_cup_demB[,c(8,22)]),"C://Users//elesa//Documents//TFG//tabla_cuponesB.csv
", row.names = FALSE)
transactions_cuponesB <-
read.transactions("C://Users//elesa//Documents//TFG//tabla_cuponesB.csv", format = "single", sep =
",", cols = c("BASKET_ID", "SUB_COMMODITY_DESC"), header =TRUE, rm.duplicates = TRUE)
soporte_cupB <- 30 / dim(transactions_cuponesB)[1]
rules_cupB <- apriori (transactions_cuponesB, parameter = list(supp = soporte_cupB, conf = 0.77))
inspect(rules_cupB)
#Mostramos la figura con las reglas obtenidas de la campaña B
plot(rules_cupB, method = "graph", engine="html")
```

Código B.8

```
#aplicamos el algoritmo a los cupones C
write.csv(na.omit(tabla_cup_demC[,c(8,22)]),"C://Users//elesa//Documents//TFG//tabla_cuponesB.csv
", row.names = FALSE)
transactions_cuponesC <-
read.transactions("C://Users//elesa//Documents//TFG//tabla_cuponesB.csv", format = "single", sep =
",", cols = c("BASKET_ID", "SUB_COMMODITY_DESC"), header =TRUE, rm.duplicates = TRUE)
soporte_cupC <- 30 / dim(transactions_cuponesC)[1]
rules_cupC <- apriori (transactions_cuponesC, parameter = list(supp = soporte_cupC, conf = 0.65))
inspect(rules_cupC)
#Mostramos la figura con las reglas obtenidas de la campaña C
plot(rules_cupC, method = "graph", engine="html")
```

8. Bibliografía

Agrawal, R., Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of International Conference on Very Large Data Bases*, pp. 487-499.

DATAtab Team (2024). DATAtab: Online Statistics Calculator. DATAtab e.U. Graz, Austria. Obtenido de https://datatab.es (Fecha de consulta: 20/10/2023)

Economipedia. (2020). Bien complementario. Obtenido de https://economipedia.com/definiciones/bien-complementario.html (Fecha de consulta: 18/10/2023)

Eurostat. (2012). Estadísticas sobre distribución de la renta. Obtenido de https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Income_distribution_statistics/es&oldid=87025#V.C3.A9ase_tambi.C3.A9n (Fecha de consulta: 06/01/2024)

Fernández Sandoval, P. (2014). Aplicación del market basket análisis para promover ofertas óptimas al cliente. *Universidad Nacional Autónoma de México*, pp. 6-7

Hahsler, M. (2017). arulesViz: Interactive visualization of association rules with R. *R Journal*, 9(2):163-175

---------- (2021). An R Companion for Introduction to Data Mining: Chapter 5. Online Book. Obtenido de https://mhahsler.github.io/Introduction_to_Data_Mining_R_Examples/book/ (Fecha de consulta: 29/11/2023)

Han, J., Kamber, M., Pei., J. (2012). Data Mining Concepts and Techniques. *Morgan Kaufmann,* pp. 245-249

La La, F. (2018). Inteligencia artificial: Análisis de la cesta de la compra. Obtenido de https://learn.microsoft.com/es-es/archive/msdn-magazine/2018/december/artificially-intelligent-market-basket-analysis (Fecha de consulta: 20/10/2023)

Márquez Sánchez, F., Sorhegui-Ortega, R. (2018). la globalización y los dilemas del desarrollo. obtenido de https://researchgate.net/publication/329428478_la_globalizacion_y_los_dilemas_del_desarrollo (fecha de consulta: 21/11/2023)

Márquez Sánchez, J. (2020). Los 10 alimentos más caros del mundo: bocados que valen su peso en oro. Obtenido de https://www.eleconomista.es/status/noticias/10703102/08/20/Los-alimentos-mas-caros-del-mundo-bocados-que-valen-su-peso-en-oro-o-algo-mas.html (Fecha de consulta: 21/12/2023)

Neves Ferraz., Cristina Bicharra Garcia, A. (2008). Ontology In Association Rules Pre-Processing And Post-Processing. *IADIS European Conference Data Mining*, pp. 87-91

Rivera, S. I. (2015). Big Data Marketing: una aproximación. *Perspectivas*, pp.147-158. ISSN 1994-3733.

Tan, P., Steinbach., M., Kumar.,V. (2006). Introduction to data mining, volume 1. *Pearson Addison Wesley Boston*

Wang, J. (2018). A Guide to Association Rules in R - Part 2 Read Market Basket Data in arules. Obtenido de https://www.jdatalab.com/data_science_and_data_mining/2018/10/15/association-rule-read-transactions.html (Fecha de consulta: 15/11/2023)

(2023). Valores faltantes. Obtenido de https://www.uv.es/webgid/Descriptiva/23_valores_faltantes.html (Fecha de consulta: 11/11/2023)