Facultad de Ingeniería

# Minería de datos

## Año: 2023

## Trabajo Integrador Final

## “E-commerce en Reino Unido”

Salim Taleb, Nasim A

Docentes: Mariana, Blanco; Juan, Aued

Carrera: Tecnicatura universitaria en procesamiento y explotación de datos

# Limpieza

Los datos provistos inicialmente constaban de 522.064 filas con 7 variables cada una, en general se observa una pobre calidad de los datos, en parte atribuida a la gran cantidad de datos nulos en algunas variables, la falta de estándar en otras, y la existencia de valores numéricos atípicos negativos y otros muy elevados, esto puede verse si se observan las filas con valores extremos en cada una de las variables.

Las operaciones de limpieza que decidieron hacerse son las siguientes:

1. Eliminar las operaciones dónde la cantidad comprada sea menor a 0, ya que representan transacciones canceladas que no aportan nada útil al análisis.
2. Eliminar operaciones dónde no esté el nombre del ítem, porque si unos de los objetivos es encontrar patrones de las compras de ítems, no tiene mucho sentido utilizar datos dónde no se sabe qué ítem se compró.
3. Eliminar operaciones dónde el precio sea negativo, por la misma razón que se filtró por cantidad, no aportan nada al análisis y no tiene sentido un precio negativo.
4. Finalmente, se observó que en algunas filas la variable de nombre del ítem tenían valores extraños, por ejemplo, “found”, “check”, “?”, etc. La gran mayoría de estos valores extraños se diferencia de las verdaderas compras, ya que contenían al menos una letra minúscula y ninguna letra mayúscula, excepto algunos valores particulares. Después de observar detenidamente los datos, y teniendo en cuenta lo anterior, se agregaron algunas palabras claves más y se eliminaron estos datos del dataset.

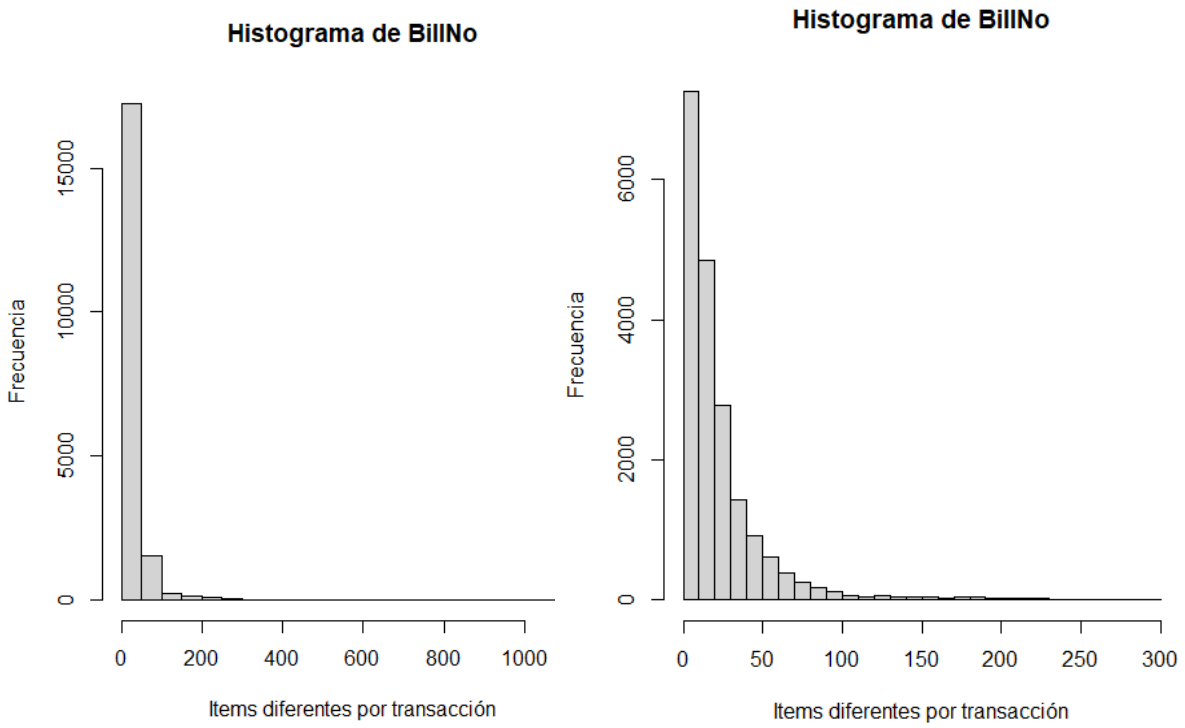
Quedaron finalmente 517.899 observaciones, es decir, que se limpiaron 4.165 datos que no aportaban nada al análisis.

# Exploración

A continuación se detalla un análisis específico de cada variable, por la gran cantidad de datos resulta difícil poder representar los datos de una manera clara en algunos casos.

## BillNo

Esta variable, si bien tiene valores numéricos, es una variable categórica que indica a qué número de compra hacer referencia la fila. Se puede visualizar su frecuencia mediante un histograma, para poder observar que cantidad de ítems únicos suelen comprarse por transacción.

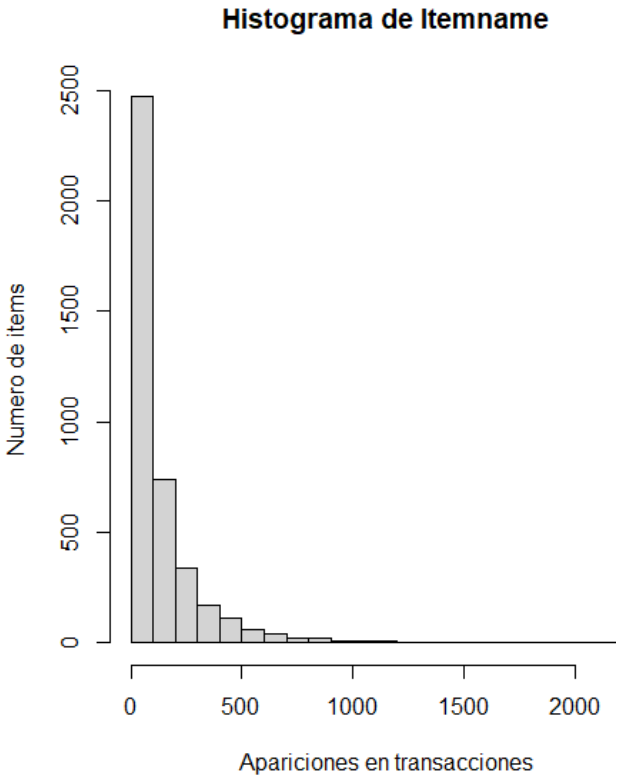



Se utilizaron dos gráficas a diferentes intervalos con el fin de visualizar los datos más cercanos a 0 que son los predominantes y que tienden a agruparse y resulta difícil diferenciarlos al tener también datos con muchos ítems por transacción.

Se ve que en la gran mayoría de transacciones el número de ítems diferentes comprados es pequeño, mayoritariamente menor a 50.

## Itemname

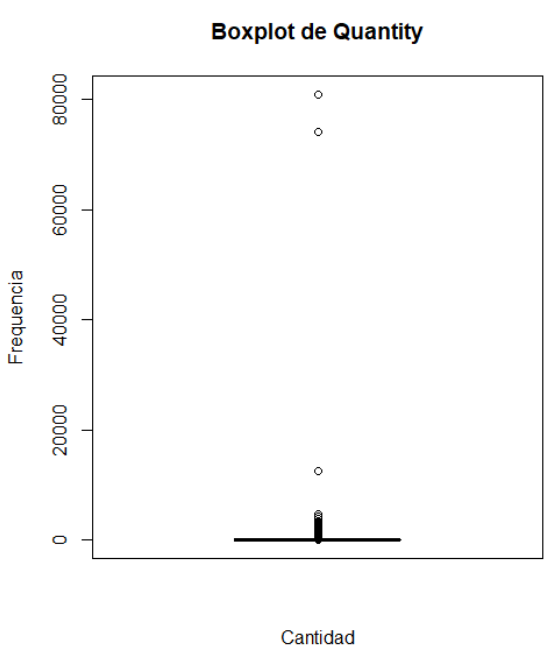
Una variable también categórica que representa al ítem en cuestión que se está comprando en la operación, se observa gran heterogeneidad en la forma de nombrar a los ítems, es decir, que los nombres no siguen ningún estándar y en algunos casos hasta presentar caracteres especiales extraños.
Para visualizar esta variable se utilizará un histograma en la cual se podrá ver la cantidad de ítems que suelen aparecer muchas veces en las compras.

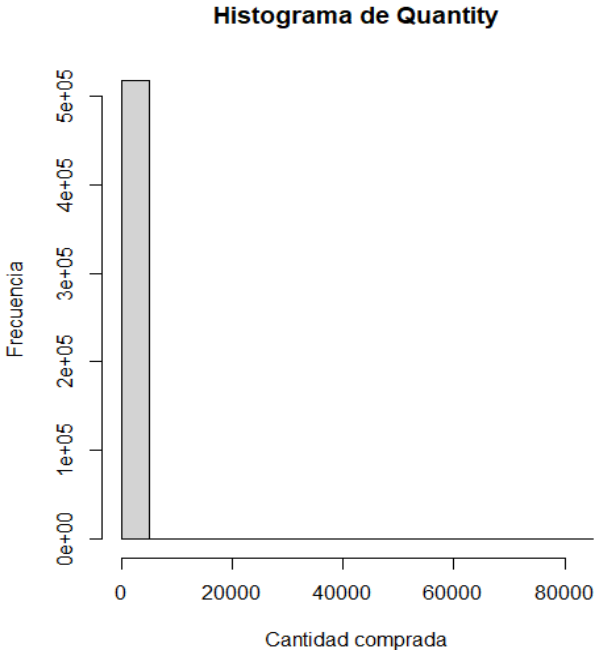



Se observa que la gran mayoría de ítems aparecen en las transacciones menos de 100 veces.

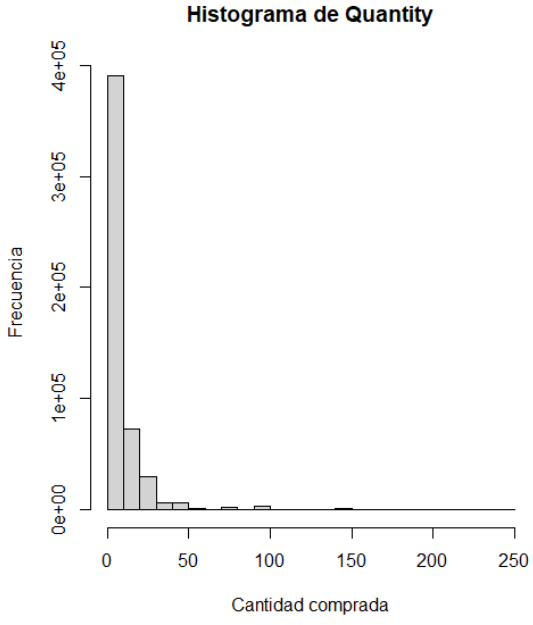
## Quantity

Si bien esta variable es numérica y puede analizarse de manera diferente a las anteriores, que eran categóricas, al ser los precios tan dispersos, no fue posible hacerlo en profundidad



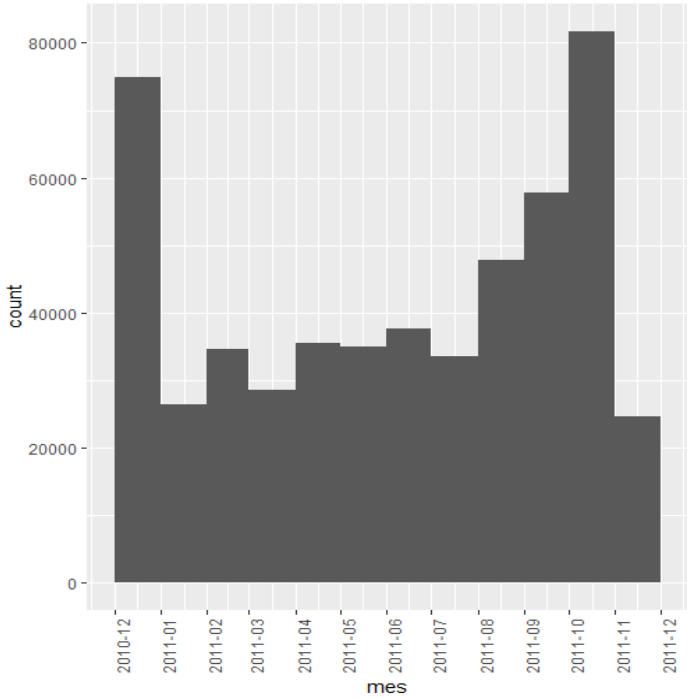

Los gráficos son intentendibles por la dispersión de los datos. Al ser que los valores muy dispersos, se hace un "zoom" imputando valores muy grandes.




Se ve finalmente que las cantidades por compra suelen ser bajos, en gran parte de las veces, menores a 10, lo que dice que por lo general no se realizan compras al por mayor.

## Date

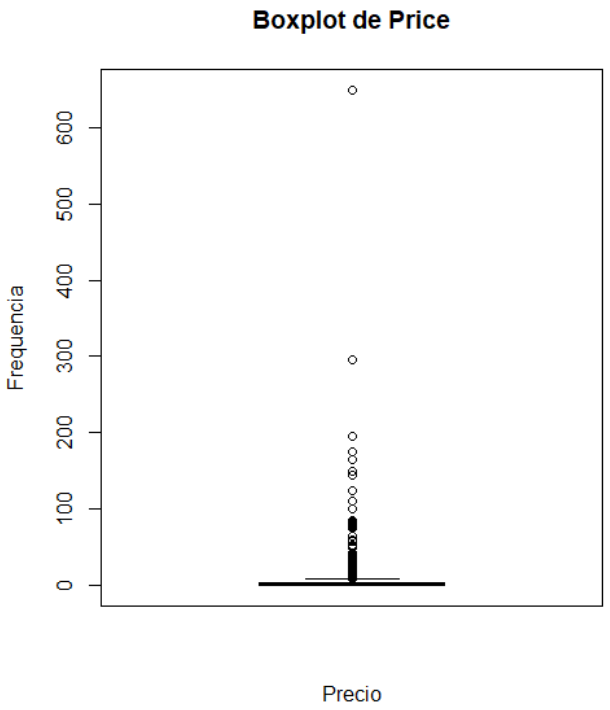
Esta variable de tipo fecha se puede explorar de manera similar a las categóricas, viendo si existe, por ejemplo, algún mes dónde hubo gran cantidad de compras, ya que visualizar los datos día por día sería imposible por la gran cantidad de diferentes días en el rango.

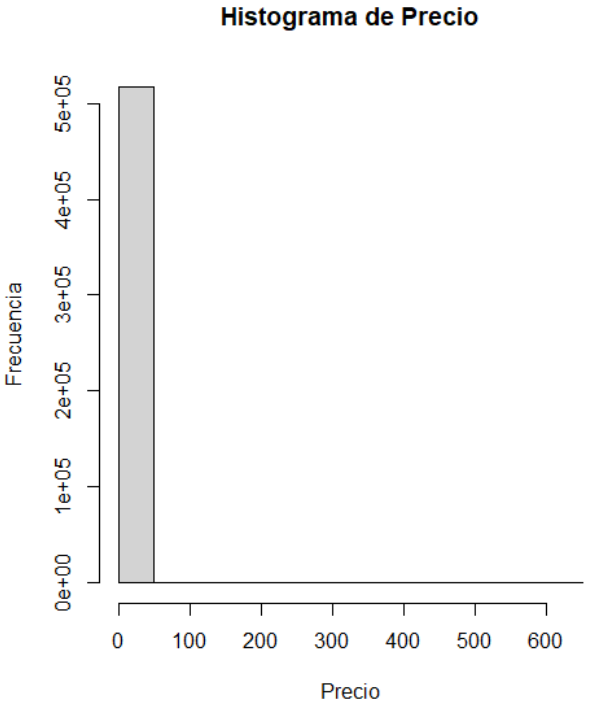


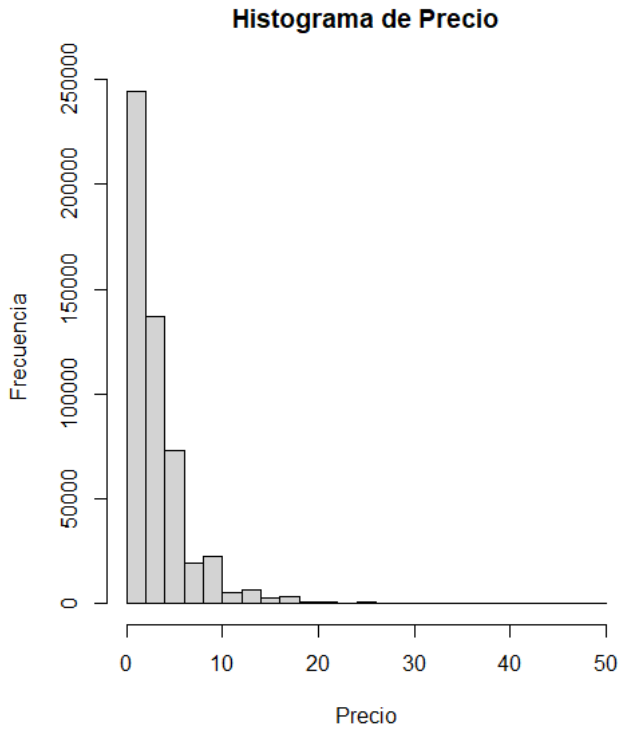

Se observa que los valores de compras únicas se mantienen más o menos constante, pero tuvieron picos durante diciembre de 2010 y la época entre agosto y noviembre de 2011.

## Price

Con esta variable ocurre algo similar a lo que pasó con “Quantity”.



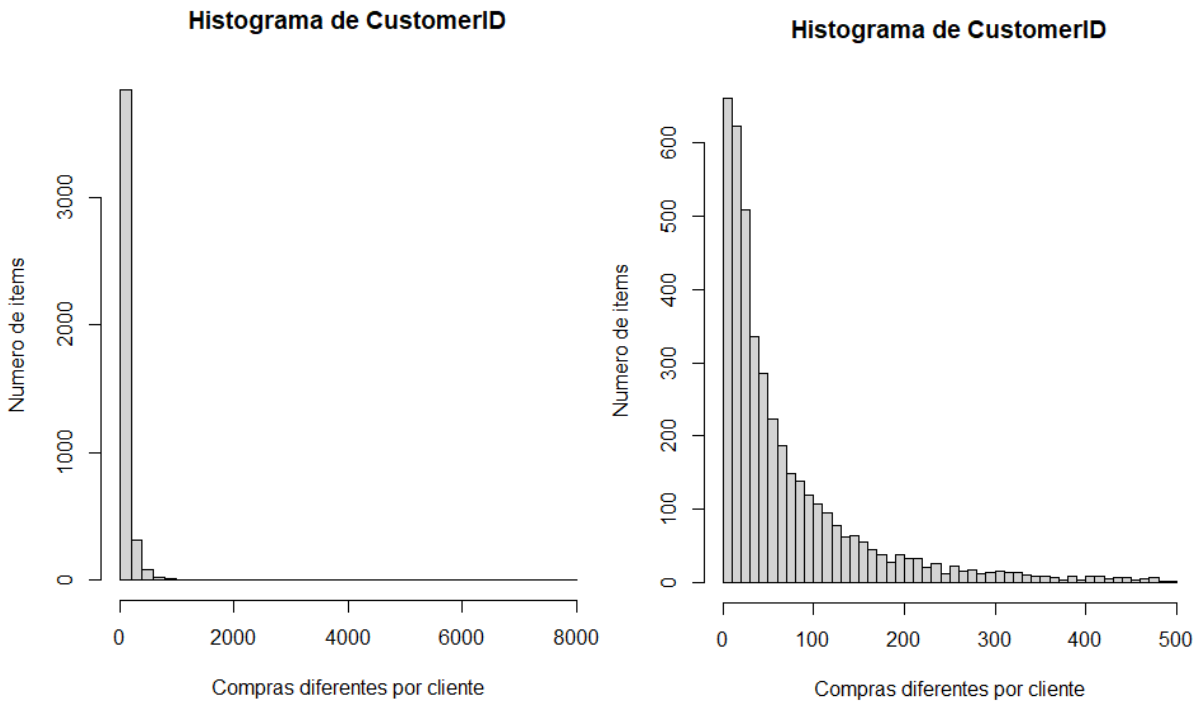




Se ve finalmente que los precios por compra suelen ser muy bajos, mayoritariamente, menores a 6, lo que indicaría que los productos vendidos suelen ser relativamente baratos.

## CustomerID

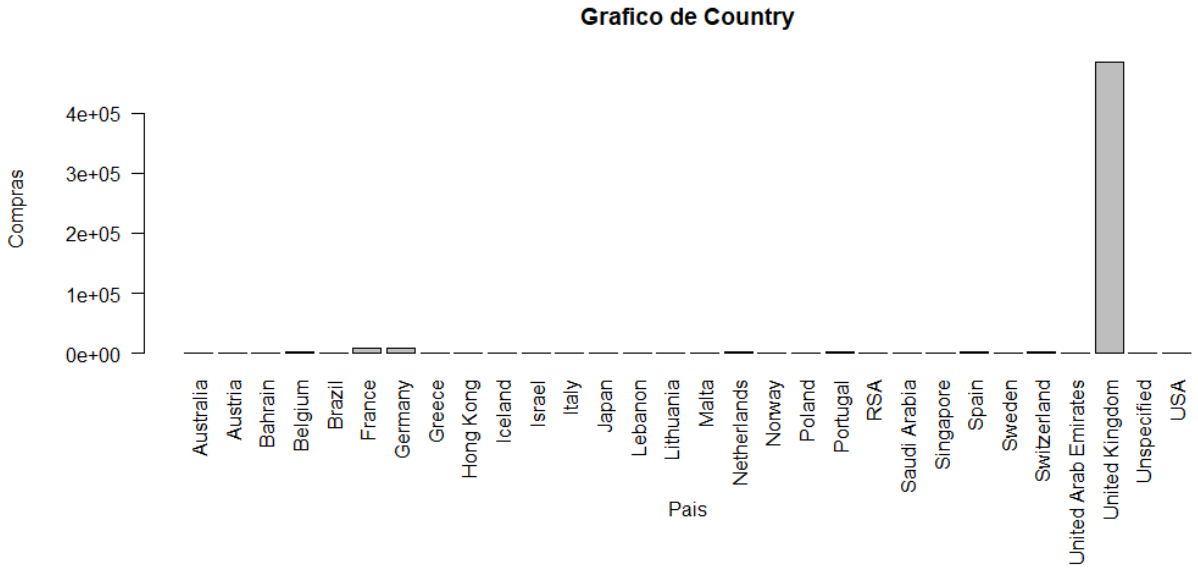
Se puede realizar un análisis similar al de las otras variables categóricas




Se observa que la gran mayoría de clientes no realiza muchas compras, la mayoría se ubica por debajo de las 200.

## Country

En este caso, al ser relativamente pocas categorías se podría utilizar un gráfico de barras.




Por lo que se ve, la gran mayoría de las compras provienen del Reino Unido, lo que tiene sentido, ya que la tienda se encuentra allí, también se ve una pequeña diferencia de ventas en países europeos a comparación con el resto, seguramente por la cercanía.

# Minado

## Secuencias

Para este paso se eliminaron las filas de compras dónde no se conocía el nombre del ítem, porque estás compras no proveen información relevante a la búsqueda de patrones, pero siguen siendo compras válidas que podían ser consideradas previamente en la exploración, por ejemplo, en las ventas por mes.

Después de agrupar las compras de ítems iguales, y quedarse con solo la primera compra de cada ítem por persona, queda un total de 18.054 transacciones y 3843 ítems.

Considerando un soporte del 2% se encontraron 4990 secuencias, las cuales se filtraron aumentando el soporte a 3% y quitando las que contenían un solo ítem, ya que no aportan información, quedando un total de 527, la mayoría con un tamaño de 2 y algunas de 3.

Las 10 de mayor soporte son las siguientes:

| Secuencia | Soporte | Tamaño |
|---|---|---|
| <{PAPER CHAIN KIT 50'S CHRISTMAS,PAPER CHAIN KIT VINTAGE CHRISTMAS}> | 0,07219375873 | 2 |
| <{GREEN REGENCY TEACUP AND SAUCER,ROSES REGENCY TEACUP AND SAUCER}> | 0,07102934327 | 2 |
| <{GREEN REGENCY TEACUP AND SAUCER,PINK REGENCY TEACUP AND SAUCER}> | 0,06474149977 | 2 |
| <{HEART OF WICKER LARGE,HEART OF WICKER SMALL}> | 0,0635770843 | 2 |
| <{RED HANGING HEART T-LIGHT HOLDER,WHITE HANGING HEART T-LIGHT HOLDER}> | 0,06217978575 | 2 |
| <{REGENCY CAKESTAND 3 TIER,ROSES REGENCY TEACUP AND SAUCER}> | 0,0605496041 | 2 |
| <{PINK REGENCY TEACUP AND SAUCER,ROSES REGENCY TEACUP AND SAUCER}> | 0,05868653936 | 2 |
| <{GARDENERS KNEELING PAD CUP OF TEA,GARDENERS KNEELING PAD KEEP CALM}> | 0,05868653936 | 2 |
| <{JUMBO BAG PINK POLKADOT,JUMBO BAG RED RETROSPOT}> | 0,05845365626 | 2 |
| <{LUNCH BAG PINK POLKADOT,LUNCH BAG RED RETROSPOT}> | 0,05798789008 | 2 |

## Ítems frecuentes

En cuanto a los 10 ítems más comprados, considerando como más comprado al que aparece en mayor cantidad de transacciones y no por cantidad, y su frecuencia relativa:

| Ítem | Frecuencia relativa |
|---|---|
| WHITE HANGING HEART T-LIGHT HOLDER | 0,1062922344 |
| REGENCY CAKESTAND 3 TIER | 0,09011853329 |
| JUMBO BAG RED RETROSPOT | 0,08712750637 |
| ASSORTED COLOUR BIRD ORNAMENT | 0,07521878808 |
| PARTY BUNTING | 0,07488645176 |
| LUNCH BAG RED RETROSPOT | 0,07012296444 |
| SET OF 3 CAKE TINS PANTRY DESIGN | 0,06159299878 |
| LUNCH BAG BLACK SKULL. | 0,0575495735 |
| PACK OF 72 RETROSPOT CAKE CASES | 0,05494627229 |
| SPOTTY BUNTING | 0,05428159965 |

## Asociaciones de compras del Reino Unido

Una vez filtradas solamente las compras desde Reino Unido y agrupadas por día, se guardan en un nuevo archivo en formato que permite utilizar la función read_baskets(), se cargan dichas compras con la función y se obtienen las siguientes reglas de asociación con un soporte mínimo del 1% y confianza del 70%, ordenadas por Lift.

| Regla | Soporte | Confianza | Lift |
|---|---|---|---|
| {HERB-MARKER-THYME} => {HERB-MARKER-ROSEMARY} | 0,01024056232 | 0,9441340782 | 86,08311985 |
| {HERB-MARKER-ROSEMARY} => {HERB-MARKER-THYME} | 0,01024056232 | 0,9337016575 | 86,08311985 |
| {HERB-MARKER-ROSEMARY} => {HERB-MARKER-PARSLEY} | 0,01005877719 | 0,9171270718 | 85,03004532 |
| {HERB-MARKER-PARSLEY} => {HERB-MARKER-ROSEMARY} | 0,01005877719 | 0,9325842697 | 85,03004532 |
| {HERB-MARKER-ROSEMARY} => {HERB-MARKER-MINT} | 0,01005877719 | 0,9171270718 | 82,25732645 |
| {HERB-MARKER-MINT} => {HERB-MARKER-ROSEMARY} | 0,01005877719 | 0,902173913 | 82,25732645 |
| {REGENCY-TEA-PLATE-ROSES} => {REGENCY-TEA-PLATE-GREEN} | 0,01163424832 | 0,7245283019 | 52,67352672 |
| {REGENCY-TEA-PLATE-GREEN} => {REGENCY-TEA-PLATE-ROSES} | 0,01163424832 | 0,845814978 | 52,67352672 |
| {POPPY'S-PLAYHOUSE-LIVINGROOM} => {POPPY'S-PLAYHOUSE-BEDROOM} | 0,01024056232 | 0,8086124402 | 51,32511962 |
| {SET-OF-3-WOODEN-TREE-DECORATIONS} => {SET-OF-3-WOODEN-STOCKING-DECORATION} | 0,01042234745 | 0,7510917031 | 49,78018625 |