UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS

FACULTAD DE INGENIERÍA

Data Mining & Data Analysis

Trabajo Final

Alumno

Hualtibamba Valerio, Mathias (U202214421)

Vilchez Marin, Rody (U202216562)

Profesor

Carlos Fernando Montoya Cubas

Lima, 03 de Diciembre de 2025

1. Introducción:

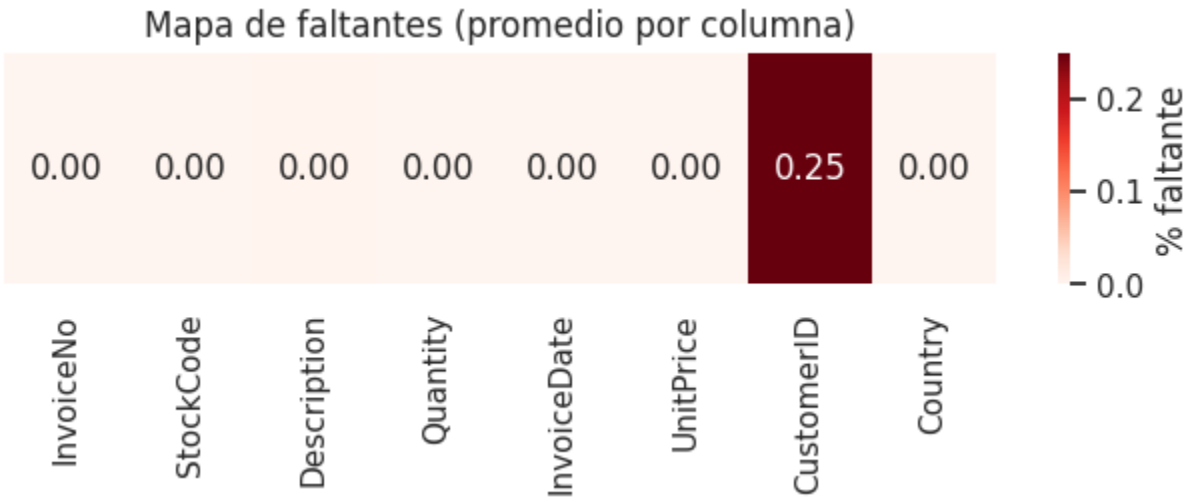
Se quiere hacer minería de patrones a un dataset transaccional con información de ventas retail mediante web durante todo el año 2011 para una tienda online proveniente de Gran Bretaña. Se busca encontrar información relevante a productos en función del negocio propuesto. El dataset contiene alrededor de 541909 campos a procesar y se dividió en dos partes tras una limpieza.

La fecha es especialmente importante en este análisis, ya que es en 2011 cuando grandes redes sociales como Instagram empezaban a nacer y otras como Facebook terminaban de asentarse y ser parte de la cultura pop; esto es importante a considerar ya que un negocio como este puede verse influenciado por modas estacionales de la época.

2. Interpretacion

2.1. EDA

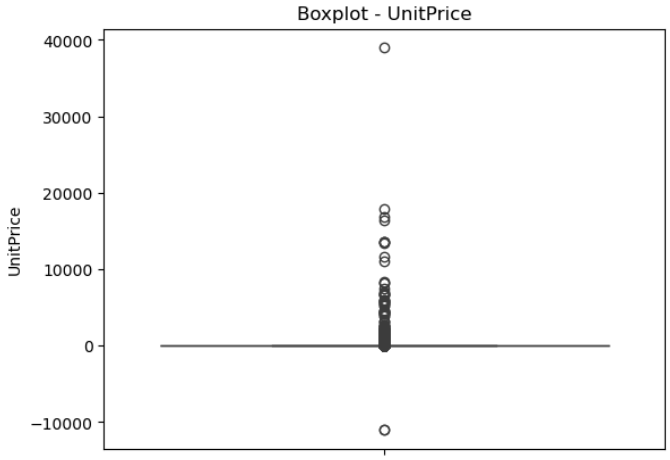
La presencia de valores faltantes finalmente se atribuyó a los usuarios que realizan compras probablemente no necesitan estar logueados. Sin embargo, el respaldo de la hipótesis fue insuficiente para sugerir mediante datos escenarios MAR o MCAR.



En cuanto a la calidad de los datos tambien se realizaron los siguientes hallazgos

```
Chequeos de calidad (conteos):
                          conteo
cancelaciones               9288
cantidades_no_positivas    10624
precios_no_positivos        2517
descripciones_vacias        1454
duplicados_exactos          5268
```

Se encontraron outliers, lo que es atribuible a errores en la página web o simples testeos por parte de los desarrolladores.



El pipeline de limpieza consto de los siguientes pasos:

1. Eliminar cancelaciones (`InvoiceNo` que empieza con "C").

2. Mantener `Quantity` y `UnitPrice` positivos.

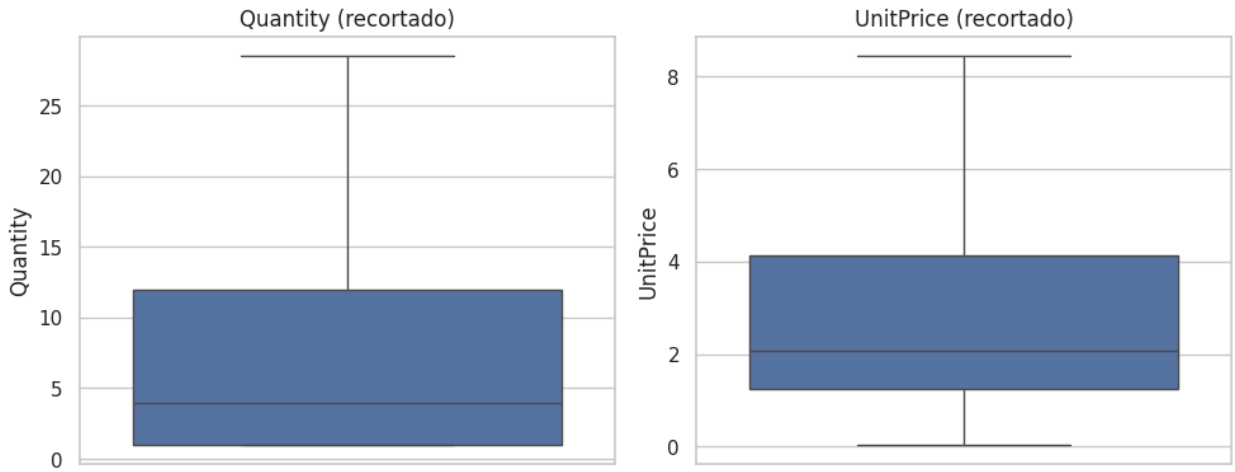
3. Quitar descripciones vacias/NA.

4. Convertir tipos y eliminar duplicados exactos.

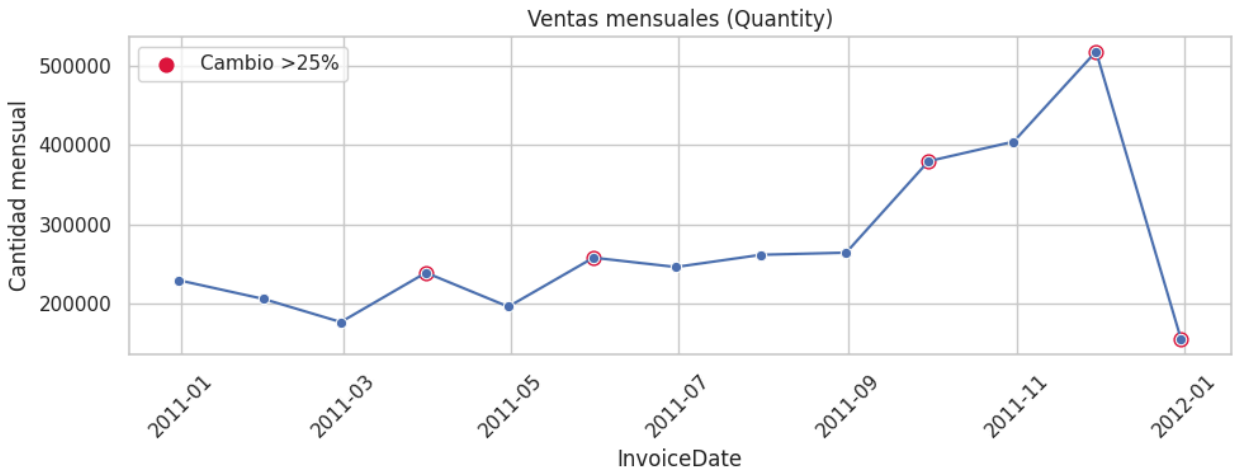
5. Filtrar productos poco frecuentes (default: >=50 apariciones) para reducir sparsity.

6. Recortar outliers via IQR (se recorta, no se elimina fila) para estabilizar distribuciones.

Quedando asi unos datos más uniformes y coherentes.



Inmediatamente a esto se quiso buscar comportamientos anómalos a lo largo del año, por lo que se hizo una gráfica resaltando picos en la variación de compras.

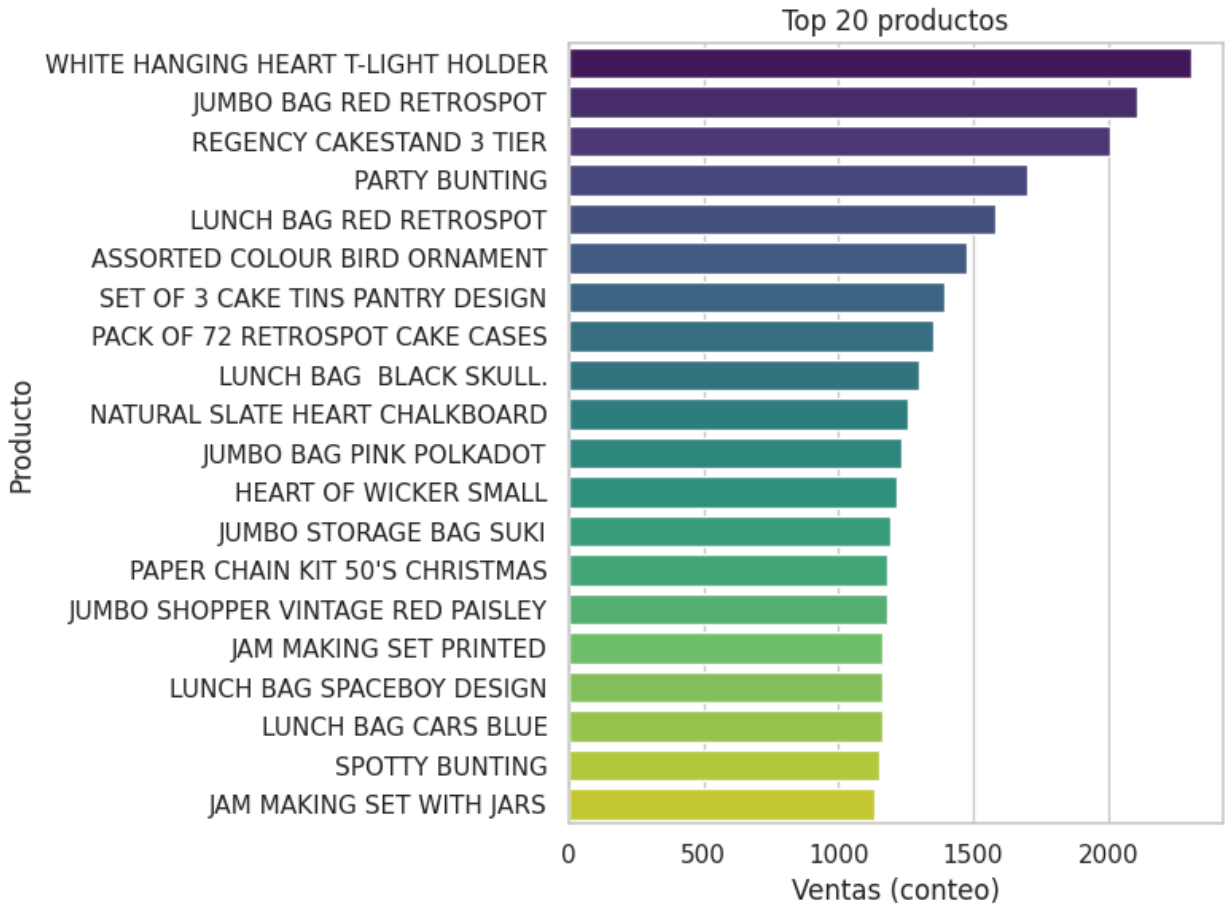


Esto nos dice algo muy curioso, y es que se puede ver claramente como las ventas se disparan desde agosto. Esto se puede deber al conjunto de fiestas y celebraciones que se realizan en esa etapa del año como Halloween, Vísperas de Navidad, Día de acción de gracias, etc.

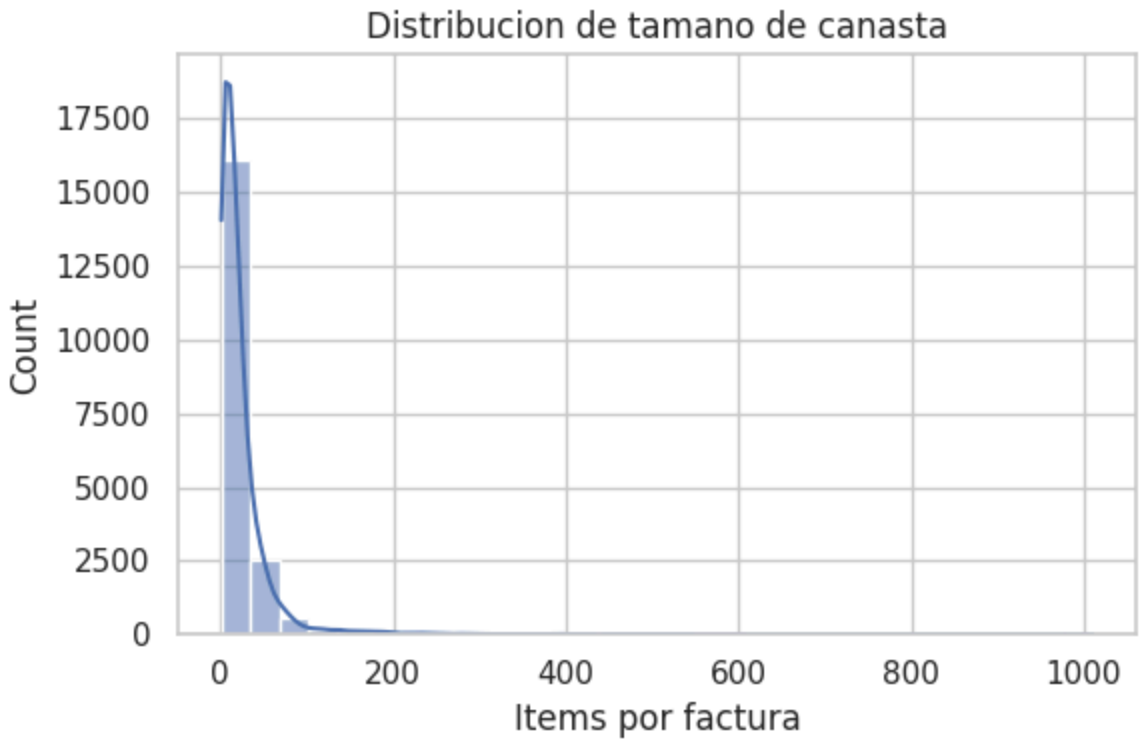
Además, podemos ver que en la primera etapa del año también hay dos picos, lo que puede atribuirse a tendencias estacionales o algún comportamiento norteamericano que no conozcamos a profundidad.

Esto nos basta para poder separar los datos en dos mitades equitativas a lo largo del tiempo.

Se hizo una gráfica de los productos más vendidos para confirmar la naturaleza del negocio y confirmar que se vende de todo.

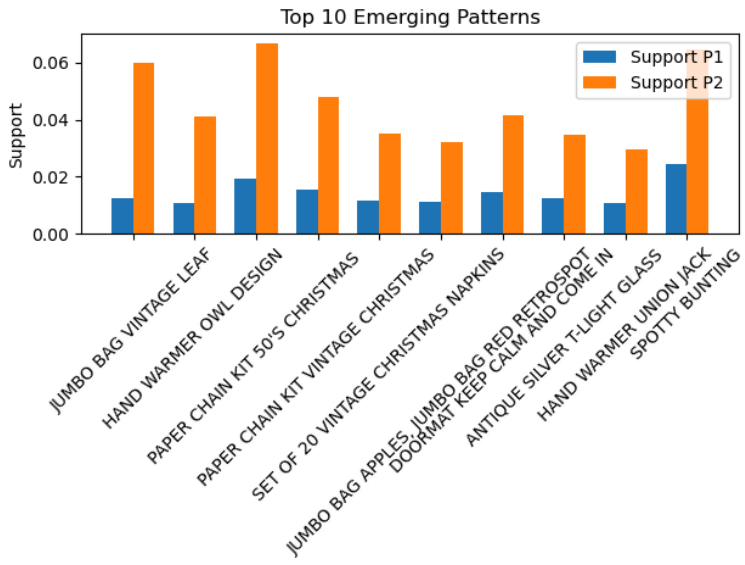


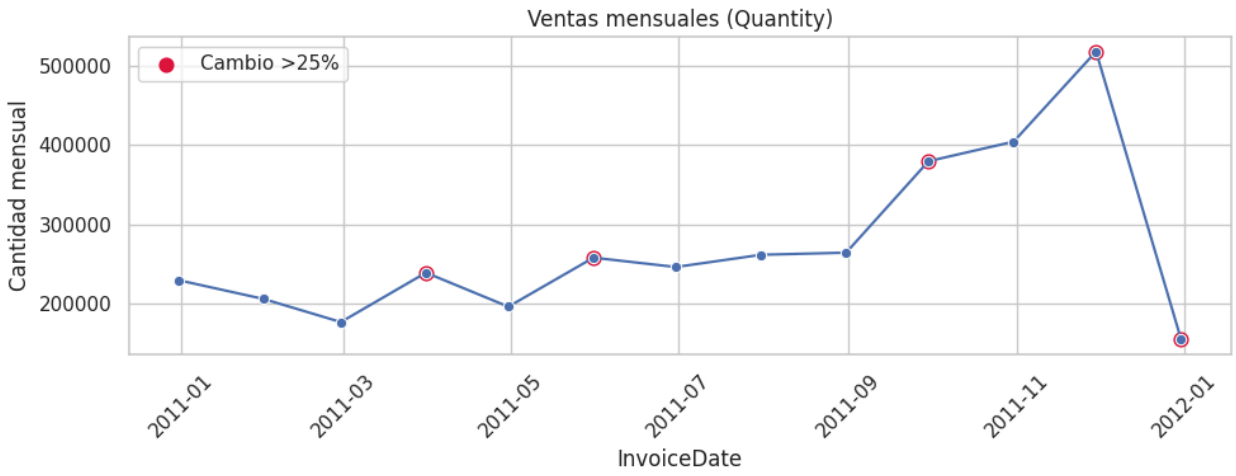
La mayor cantidad de gente compra 15 articulos, pero hay quien llega a realizar compras en linea de cifras cercanas a los 1000 articulos.



2.2. Pattern Mining

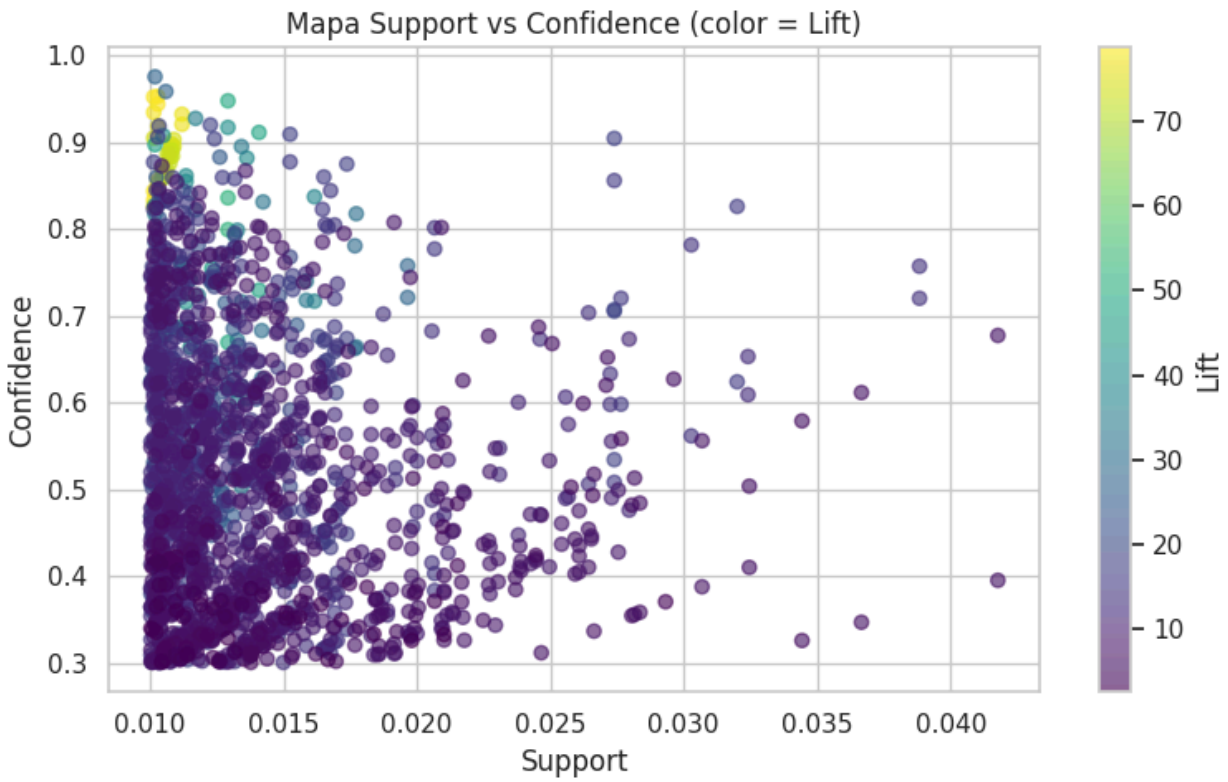
Se intentó identificar la variación en la venta de productos. Se obtuvieron, a través de las siguientes gráficas, resultados que confirman la naturaleza esperada del negocio:





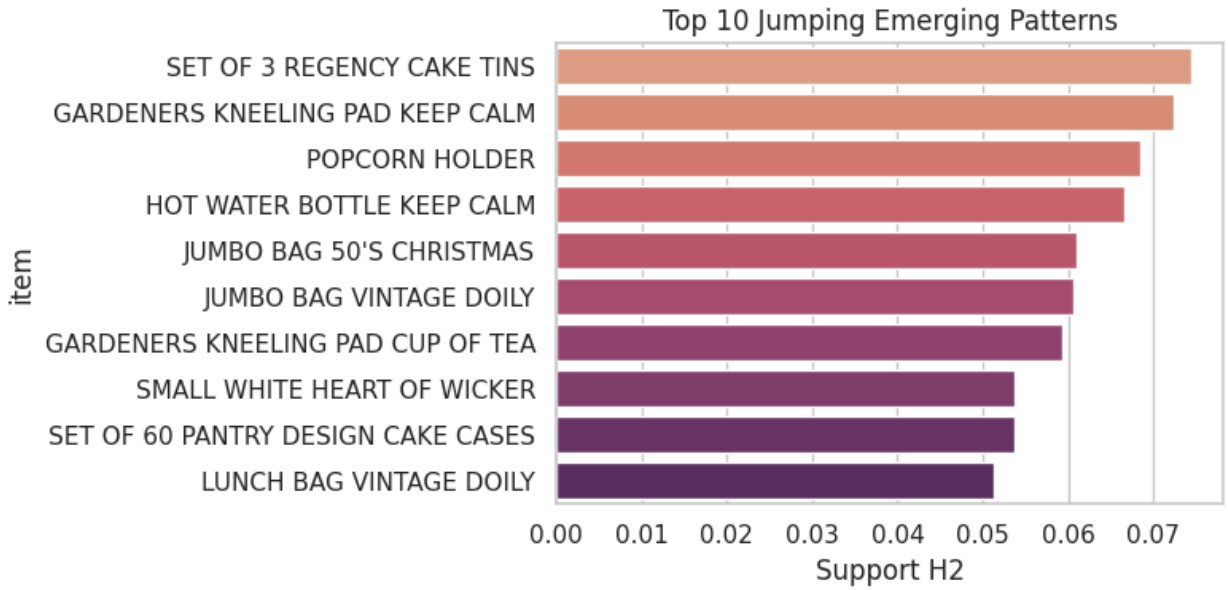
Esto nos dice que los productos navideños fueron los que resultaron en un aumento considerable, triplicando incluso sus ventas. Además, productos sin razón alguna como SPOTTY BUNTING también crecieron. Estos comportamientos que parecen tan estacionales y a la vez algo aleatorios solo nos dice que el negocio como tal, naturalmente, no tiene una identidad propia. Esto no es malo per se, sin embargo nos dice que tiene la facultad de poder adaptarse de mejor manera a los cambios. Es decir, el negocio debe estar atento a las tendencias y modas para empezar a promover estos productos, así como dejar de lado aquellos estacionales como productos navideños ya que el negocio de por sí será concurrido y se venderán esos items por la misma naturaleza que tienen.

Reglas de asociacion



	consequents	support	confidence	lift
1696	(HERB MARKER THYME)	0.0103	0.9442	78.6742
1699	(HERB MARKER PARSLEY, HERB MARKER ROSEMARY)	0.0103	0.8565	78.6742
1700	(HERB MARKER PARSLEY, HERB MARKER THYME)	0.0103	0.8458	78.4203
1695	(HERB MARKER ROSEMARY)	0.0103	0.9531	78.4203
1795	(HERB MARKER THYME, HERB MARKER BASIL)	0.0101	0.8333	78.3651

El análisis de reglas de asociación muestra un grupo de hierbas (parsley, rosemary, thyme y basil) que aparece casi siempre en conjunto, con confidences superiores al 85% y lifts cercanos a 78, valores excepcionalmente altos. Esto indica una dependencia casi determinística entre estos ítems y revela un patrón estructural más que casual: probablemente corresponden a productos empaquetados, registros simultáneos o un patrón de uso altamente consistente. La interpretación general es que estos marcadores no actúan como compras independientes, sino como un bloque funcional que se comporta como una unidad dentro del dataset.



Los resultados muestran un volumen significativo de patrones emergentes entre H1 y H2: de los 4,135 itemsets comparados, 1,966 califican como Jumping Emerging Patterns (JEP) al presentar soporte cero en H1 y aparición en H2. Los más destacados —SET OF 3 REGENCY CAKE TINS, GARDENERS KNEELING PAD KEEP CALM, POPCORN HOLDER, HOT WATER BOTTLE KEEP CALM y JUMBO BAG 50'S CHRISTMAS— exhiben soportes entre 6% y 7% en H2, con growth rate infinito, lo que indica una aparición súbita y contundente. En conjunto, este comportamiento sugiere cambios estructurales en la oferta o en la demanda: incorporación de nuevos productos, campañas estacionales o eventos comerciales que no estaban presentes en H1 y que ahora generan un impacto notable en las transacciones.

Conclusiones

El análisis exploratorio confirmó que el comportamiento de compra en este e-commerce británico está fuertemente condicionado por la estacionalidad del segundo semestre, donde eventos como Halloween y la temporada navideña generan un aumento abrupto en la actividad transaccional. La limpieza de datos permitió estabilizar el dataset y evidenció tanto errores operativos (outliers, cancelaciones, duplicados) como la presencia natural de clientes no registrados.

El análisis de variación temporal mostró que numerosos productos incrementan su demanda exclusivamente en ventanas específicas del año. Esto se reforzó mediante *pattern mining*: más del 47% de los itemsets comparados calificaron como **Jumping Emerging Patterns**, apareciendo por primera vez en H2 con soportes elevados, señalando una transición clara en la estructura de la demanda.

El análisis de reglas de asociación mostró patrones fuertemente determinísticos, como el cluster de hierbas (parsley, rosemary, thyme, basil), su co-ocurrencia casi perfecta sugiere productos empaquetados o registros estructurales, más que preferencias libres del consumidor. Esto indica que no todos los patrones reflejan comportamiento natural del cliente; algunos representan

En conjunto, los hallazgos muestran que este negocio no mantiene una identidad de consumo estable, sino que opera en un entorno altamente modulable, con ciclos de demanda dependientes de tendencias, campañas y temporadas. Su fortaleza radica precisamente en esa capacidad de adaptación: productos emergen, se consolidan y desaparecen según el contexto temporal, la oferta y la moda del consumidor digital en 2011.

Recomendaciones

- Planificar inventario de forma estacional (no uniforme).
  Las curvas detectadas confirman que el stock debe incrementarse estratégicamente en el segundo semestre, particularmente en categorías navideñas y decorativas, mientras que el primer semestre puede operarse con inventario más conservador.

- Detectar y priorizar JEPs como señales tempranas de tendencia.
  Los productos con soporte cero en H1 y alto soporte en H2 deben tratarse como oportunidades: su aparición súbita indica capacidad viral, moda o cambio de catálogo. Se recomienda implementar monitoreo trimestral de *emerging patterns* para reaccionar antes que la competencia.

- Construir bundles y estrategias de cross-selling basadas en clusters determinísticos.
  Los itemsets con lift extremo (como el set de hierbas) justifican crear paquetes, descuentos combinados o recomendaciones automáticas, ya que los clientes los adquieren juntos de forma consistente.

- Depurar procesos operativos para reducir outliers y errores transaccionales.
  Los outliers identificados sugieren problemas de registro, pruebas internas o fallos del sistema. Se recomienda implementar validaciones en frontend y backend que limiten cantidades no plausibles y capturen inconsistencias antes de ingresar al sistema.

- Refinar campañas de marketing basadas en ciclos detectados.
  La variación temporal indica que campañas genéricas son menos efectivas. Conviene orientar anuncios, banners y promociones a eventos estacionales concretos (Halloween, Navidad, etc.), incorporando productos emergentes y retirando progresivamente aquellos cuyo soporte decae.