

Online Retail II EDA

Índice

1. Introducción	3
1.1. Motivación	3
1.2. Justificación	3
1.3. Objeto de estudio	3
1.4. Problema de investigación	3
1.4.1. Problema general	3
1.5. Objetivos	4
1.5.1. Objetivo general	4
1.5.2. Objetivos específicos	4
1.6. Hipótesis	4
1.6.1. Hipótesis específicas	4
2. Descripción General del Dataset	5
2.1. ¿Podemos describir qué es un registro?	5
2.2. ¿Cuántos registros existen?	5
2.3. ¿Son pocos o demasiados registros?	5
2.4. ¿Qué representa cada fila?	5
2.5. ¿Cuáles son los tipos de datos de cada columna?	6
2.6. ¿Es una data etiquetada?, ¿cómo se interpreta la información de las clases?	6
2.7. ¿Hay niveles de granularidad de los datos?	7
2.8. ¿Los datos tienen diferentes unidades de medida?	7
3. Limpieza de los datos	8
3.1. ¿Existen datos duplicados?	8
3.2. ¿Están todas las filas completas o tenemos campos con valores nulos?	9
3.3. En caso que haya demasiados nulos: ¿Queda el resto de información útil?	9
3.4. Hipótesis 1 - Concatenación de CustomerID	10
3.5. Acciones frente a los datos nulos	15
3.6. ¿Todos los datos están en su formato adecuado?	16
3.7. Análisis de outliers	20
3.7.1. ¿Cuáles son los Outliers?	20
3.7.2. ¿Podemos eliminarlos? ¿Es importante conservarlos?	22
4. Exploración	23
4.1. ¿Cuántos productos, clientes y países existen?	23
4.2. ¿Cuáles son los productos más vendidos?	23
4.3. ¿Cuáles son los productos más vendidos por cantidad?	24
4.4. ¿Siguen alguna distribución?	24
4.5. ¿Entre qué rangos están los datos?	25
4.6. ¿Cuál es la tendencia diaria de ventas?	30
4.7. ¿Cuál es la tendencia mensual de ventas?	32
4.8. ¿Cuál es el número de transacciones por horas?	34
4.9. ¿Cuáles son los productos más vendidos por país?	36
4.10. ¿Cuáles son los clientes con mayor y menor monto total de compra?	37

4.11. ¿Cómo se distribuyen las compras totales por país?	39
4.12. ¿Cómo evoluciona la retención de clientes a lo largo del tiempo?	39
4.13. ¿Cuál es la cantidad promedio de productos comprados por cohorte trimestral?	40
4.14. ¿Existe correlación entre cantidad, precio unitario y monto total de compra?	42
4.15. ¿Cómo se relaciona la cantidad comprada con el monto total según el país de origen?	43
4.16. Hipótesis 2 - Anomalías Geográficas en Patrones de Compra	46

1. Introducción

1.1. Motivación

El comercio electrónico ha revolucionado la relación entre empresas y consumidores, generando volúmenes masivos de datos transaccionales cada día. Sin embargo, los datos por sí solos no generan valor: es su análisis inteligente lo que permite descubrir insights accionables.

Este proyecto nace del interés por aplicar técnicas avanzadas de inteligencia de negocios y minería de datos en un escenario real y desafiante.

El dataset ***Online Retail II*** ofrece la oportunidad única de trabajar con información transaccional auténtica que refleja comportamientos reales de compra, permitiéndome desarrollar competencias analíticas mientras genero conocimiento aplicable a cualquier negocio de e-commerce que busque mejorar su competitividad y retención de clientes.

1.2. Justificación

Este proyecto se justifica por su triple impacto:

- **Valor empresarial:** La segmentación efectiva de clientes mejora la retención y efectividad del marketing, identificando patrones de compra que anticipan necesidades y optimizan estrategias de ventas complementarias.
- **Rigor metodológico:** El proyecto aborda desafíos reales como el desbalance geográfico, valores atípicos y datos faltantes, evidenciando capacidad para trabajar con datos imperfectos propios de entornos empresariales.
- **Aplicabilidad inmediata:** Los resultados se traducen en estrategias concretas como campañas segmentadas, programas de fidelización personalizados y recomendaciones de productos, desarrollando un marco analítico replicable que apoya la toma de decisiones basada en evidencia, no en intuición.

1.3. Objeto de estudio

El objeto de estudio son las transacciones comerciales contenidas en el dataset ***Online Retail II***, que registran las compras realizadas por clientes entre 2009 y 2011, vinculando información de clientes, facturas y productos adquiridos.

1.4. Problema de investigación

1.4.1. Problema general

¿Qué patrones de comportamiento de compra y segmentos de clientes pueden identificarse en el dataset ***Online Retail II*** para diseñar estrategias efectivas de fidelización, considerando los desafíos de calidad de datos como desbalance geográfico, valores atípicos y datos faltantes?

La identificación de estos patrones es fundamental para la gestión efectiva de relaciones con clientes en el comercio electrónico [6, 2].

1.5. Objetivos

1.5.1. Objetivo general

Identificar patrones de comportamiento de compra en el dataset *Online Retail II* para apoyar la toma de decisiones estratégicas orientadas a la fidelización de clientes, aplicando técnicas de minería de datos y análisis de comportamiento del consumidor [1, 4].

1.5.2. Objetivos específicos

1. Preparar el dataset *Online Retail II* garantizando la calidad de los datos. [4].
2. Caracterizar el perfil de comportamiento de compra de los clientes. [1].
3. Identificar patrones de asociación entre productos en las transacciones de compra. [2].
4. Identificar grupos de clientes con comportamientos de compra similares. [7, 6].
5. Determinar los factores de comportamiento asociados a la retención de clientes. [3, 8].

1.6. Hipótesis

A partir del análisis exploratorio de datos y los desafíos identificados en el dataset *Online Retail II*, se plantean las siguientes hipótesis específicas que guiarán el proceso de preparación y análisis de datos:

1.6.1. Hipótesis específicas

1. Hipótesis 1 - Concatenación de CustomerID:

Existen CustomerID concatenados accidentalmente en el campo Description del dataset, lo cual explica parcialmente la presencia de valores nulos en la columna CustomerID.

2. Hipótesis 2 - Anomalías y relación geográfica:

Existen datos anómalos en las variables Quantity y UnitPrice que están asociados sistemáticamente con países específicos, indicando diferencias en los patrones de compra entre mercados geográficos (mayoristas vs minoristas, productos de lujo vs estándar).

3. Hipótesis 3 - Patrones temporales de compra:

Las transacciones de compra presentan patrones temporales identificables (cíclicos, estacionales o de tendencia) que permiten anticipar períodos de alta demanda y diseñar estrategias de inventario y marketing diferenciadas.

2. Descripción General del Dataset

2.1. ¿Podemos describir qué es un registro?

En el dataset *Online Retail II*, un registro es una transacción individual, es decir, la venta de un producto específico asociada a un detalle de factura.

2.2. ¿Cuántos registros existen?

El dataset *Online Retail II* cuenta con un total de **1 067 371 registros**, cada uno de los cuales representa una transacción individual de un producto dentro de un detalle de factura.

2.3. ¿Son pocos o demasiados registros?

La cantidad de registros es considerablemente alta. Se trata de un volumen de datos lo suficientemente grande como para llevar a cabo análisis exploratorios robustos, segmentaciones de clientes, detección de patrones de consumo y estudios de comportamiento de compra.

Este tamaño permite obtener resultados estadísticamente significativos y aplicar técnicas avanzadas de análisis, como minería de datos o aprendizaje automático, con una base de información confiable.

2.4. ¿Qué representa cada fila?

En el dataset *Online Retail II*, cada fila representa una transacción individual de un producto dentro de una factura. Es decir, un registro equivale a una **línea de detalle de factura**.

La Tabla 1 resume los principales atributos de cada fila.

Atributo	Tipo de dato	Descripción
InvoiceNo	Object	Identificador único de la factura. Una factura puede contener varias filas.
StockCode	Object	Código único asignado a cada producto.
Description	Object	Nombre o descripción del producto vendido.
Quantity	Entero	Número de unidades vendidas del producto en la transacción.
InvoiceDate	Object	Fecha y hora exacta en que se emitió la factura.
UnitPrice	Float	Precio unitario del producto (en libras esterlinas).
CustomerID	Float	Identificador único del cliente que realizó la compra.
Country	Object	País desde el cual se efectuó la compra.

Tabla 1: Atributos que conforman cada fila (registro) en el dataset Online Retail.

De manera esquemática, la entidad relación de los datos puede visualizarse en la Figura 1, donde, un cliente tiene un país, un cliente tiene una a muchas detalles de facturas y un datalle de factura tiene un producto).

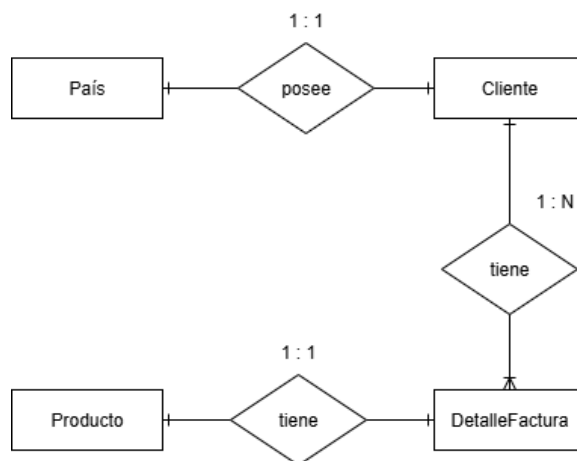


Figura 1: Diagrama entidad relación en *Online Retail II*.

2.5. ¿Cuáles son los tipos de datos de cada columna?

Atributo	Tipo de dato
InvoiceNo	Object
StockCode	Object
Description	Object
Quantity	Entero (int64)
InvoiceDate	Object
UnitPrice	Decimal (float64)
CustomerID	Decimal (float64)
Country	Object

Tabla 2: Tipos de datos de cada columna en *Online Retail II*.

2.6. ¿Es una data etiquetada?, ¿cómo se interpreta la información de las clases?

El dataset **no es una data etiquetada** en el sentido estricto de aprendizaje supervisado, ya que no incluye una variable objetivo que clasifique directamente los registros.

No obstante, es posible derivar etiquetas a partir de los datos para futuros análisis o modelos predictivos. Algunos ejemplos de posibles clases serían:

- Clasificación de clientes según su país de origen (**Country**).
- Clasificación de productos según códigos (**StockCode**) o descripciones (**Description**).

2.7. ¿Hay niveles de granularidad de los datos?

Sí, el dataset presenta diferentes niveles de granularidad que permiten analizar la información desde una vista general hasta el detalle más específico:

- **Geográfico:** A nivel de país (`Country`).
- **Cliente:** A nivel de cliente individual (`CustomerID`).
- **Factura:** A nivel de transacción agrupada (`InvoiceNo`).
- **Producto:** A nivel de línea de detalle de factura (`StockCode`, `Description`, `Quantity`, `UnitPrice`).
- **Temporal:** A nivel de fecha y hora exacta (`InvoiceDate`), lo cual permite análisis por año, mes, día, hora o minuto.

Estos distintos niveles permiten realizar análisis tanto agregados (por países o periodos de tiempo) como detallados (compras específicas por cliente y producto en un momento dado).

2.8. ¿Los datos tienen diferentes unidades de medida?

Los datos numéricos están en unidades consistentes:

- **Quantity:** número de unidades vendidas.
- **UnitPrice:** precio en libras esterlinas (£).

No se observan múltiples unidades de medida en un mismo campo.

3. Limpieza de los datos

3.1. ¿Existen datos duplicados?

Durante el análisis preliminar se identificó la presencia de registros duplicados en el dataset *Online Retail*. Estos duplicados pueden deberse a errores de carga o a repeticiones innecesarias en las transacciones. Por tal motivo, resulta necesario aplicar un proceso de limpieza que elimine dichas redundancias mediante herramientas como `drop_duplicates()` en *pandas*, con el fin de garantizar la calidad de los datos y evitar sesgos en los análisis posteriores.

Para la detección de duplicados se emplearon las funciones `mostrarDuplicados()` y `mapaDuplicadosTodas()`. Los resultados obtenidos se muestran en la Tabla 3, mientras que la Figura 2 ilustra la distribución de estos registros en las distintas columnas del dataset.

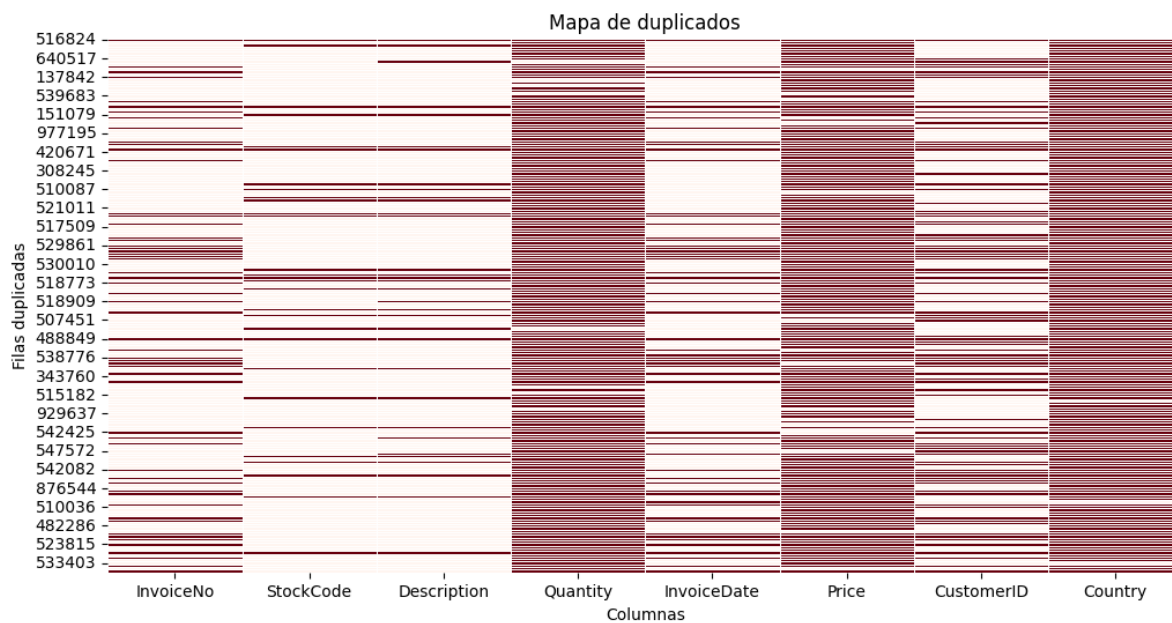


Figura 2: Mapa de calor de los registros duplicados en el dataset *Online Retail*.

Dataset	Cantidad de filas duplicadas
<i>Online Retail II</i>	34,335

Tabla 3: Cantidad de registros duplicados detectados en el dataset *Online Retail II*.

Posteriormente, se eliminaron dichos registros duplicados.

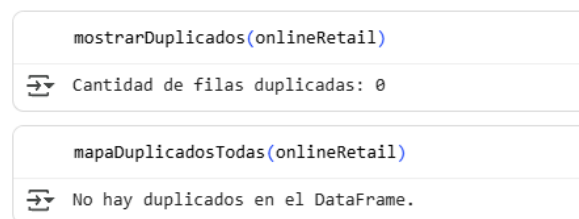


Figura 3: Resultados de la limpieza de duplicados en el dataset *Online Retail II*.

3.2. ¿Están todas las filas completas o tenemos campos con valores nulos?

Durante la verificación de valores nulos en el dataset *Online Retail*, se identificó que algunas columnas presentan registros incompletos. En particular, la columna **CustomerID** contiene **243,007 valores faltantes** y la columna **Description** presenta **4,382 valores faltantes**, mientras que las demás columnas se encuentran completas. La Tabla 4 y la figura 4 resumen los resultados obtenidos.



Figura 4: Mapa de calor de los valores nulos en el dataset *Online Retail* antes de la limpieza.

Columna	Número de valores nulos
InvoiceNo	0
StockCode	0
Description	4,382
Quantity	0
InvoiceDate	0
UnitPrice	0
CustomerID	243,007
Country	0

Tabla 4: Número de valores nulos por columna en el dataset *Online Retail*.

3.3. En caso que haya demasiados nulos: ¿Queda el resto de información útil?

Aunque la columna **CustomerID** tiene una cantidad significativa de valores faltantes, el resto de los atributos de las transacciones se encuentra mayormente completo. Por lo tanto, la información no resulta inútil; sin embargo, la ausencia del identificador limita los análisis centrados en el comportamiento individual de los clientes.

3.4. Hipótesis 1 - Concatenación de CustomerID

Contexto

Durante la exploración inicial del dataset *Online Retail II*, se identificaron **235,151 transacciones sin CustomerID registrado**. Esta cantidad considerable motivó el planteamiento de las siguientes hipótesis:

En la Figura 5 se observa la comparación entre los registros que poseen un **CustomerID** nulo frente a los que contienen un valor válido. Este análisis inicial evidencia la magnitud del problema y justifica un examen más profundo sobre el patrón de los identificadores faltantes.

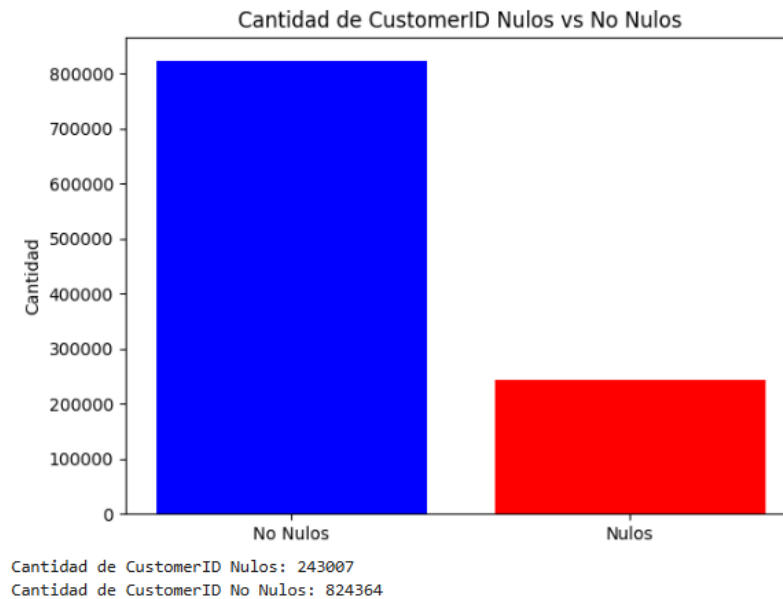


Figura 5: Comparación entre registros con CustomerID nulo y no nulo.

Posteriormente, se analizó la distribución del número de dígitos presentes en los identificadores válidos. La Figura 6 muestra que el promedio se concentra en cinco dígitos, lo cual reforzaba la expectativa de que los números incrustados en la descripción pudieran estar representando un CustomerID.

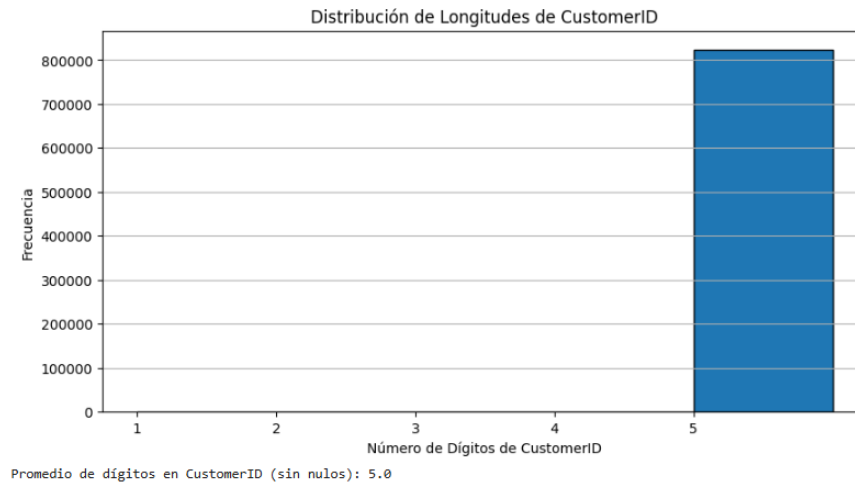


Figura 6: Distribución promedio de dígitos en los CustomerID.

Formulación de hipótesis

- H_0 (**Hipótesis nula**): Los valores nulos en CustomerID NO son producto de concatenación accidental con el campo Description. Las secuencias numéricas de 5 dígitos al final de Description son códigos de producto u otros identificadores no relacionados con CustomerID.
- H_1 (**Hipótesis alternativa**): Los valores nulos en CustomerID SÍ son producto de concatenación accidental con el campo Description. Las secuencias numéricas de 5 dígitos al final de Description corresponden a CustomerID válidos que fueron concatenados erróneamente durante el proceso de carga de datos.

Predicciones verificables

Si H_1 es correcta, deberíamos observar:

1. Filas con CustomerID nulo contienen secuencias numéricas de 5 dígitos en Description
2. Esos números coinciden con CustomerID existentes en otras transacciones del dataset
3. El patrón es exclusivo (o predominante) en filas con CustomerID nulo
4. La cantidad de coincidencias es significativa estadísticamente (¿5 % de los nulos)

Metodología

El análisis se desarrolló en 5 pasos:

- **Paso 1:** Separación de datasets (nulos vs válidos)
- **Paso 2:** Búsqueda de patrón regex en descripciones
- **Paso 3:** Verificación cruzada con CustomerID reales
- **Paso 4:** Análisis de control (grupo de comparación)
- **Paso 5:** Visualización y conclusión basada en evidencia

PASO 1: Separación de Datasets Primero se dividió el dataset en dos grupos:

- **Filas nulas:** Transacciones donde `CustomerID` es `NaN` ($n = 235,151$)
- **Filas válidas:** Transacciones con `CustomerID` registrado ($n = 797,885$)

Esta separación permitió:

1. Analizar las características específicas de cada grupo
2. Buscar el patrón sospechoso solo en el grupo relevante (nulas)
3. Usar el grupo válido como control para validar los hallazgos

PASO 2: Búsqueda de Patrón Se utilizaron expresiones regulares (regex) para buscar secuencias de **exactamente 5 dígitos en cualquier posición** de la columna `Description`.

Patrón regex utilizado: `r'\b(\d{5})\b'`

- `\b` = límite de palabra (evita coincidencias dentro de números más largos)
- `\d{5}` = exactamente 5 dígitos consecutivos
- `\b` = límite de palabra al final (asegura que sean exactamente 5 dígitos)

Ejemplos de coincidencia:

- ✓ `'1733 mixed 21733'` → extrae 21733
- ✓ `'invcd as 84879?'` → extrae 84879
- ✓ `'sold as 17003?'` → extrae 17003
- ✓ `'wrong barcode (22467)'` → extrae 22467
- × `'PACK 123456 ITEMS'` → no coincide (6 dígitos)

Resultados:

Métrica	Valor
Filas con patrón de 5 dígitos	240
Filas sin patrón	234,911
Cobertura	0.10 %

Tabla 5: Resultados de la búsqueda de patrón - Paso 2

PASO 3: Verificación Cruzada Se verificó si los números de 5 dígitos extraídos corresponden a **CustomerID reales** que aparecen en otras transacciones del dataset.

Proceso:

1. Extraer todos los `CustomerID` únicos del grupo válido
2. Convertir los números extraídos a enteros
3. Comparar ambos conjuntos (intersección)
4. Calcular tasa de coincidencia

Resultados:

Métrica	Valor
CustomerID únicos en el dataset	5,942
Números extraídos de Description	240
Coincidencias con CustomerID reales	219
Tasa de coincidencia	91.25 %
Cobertura sobre nulos	0.0931 %

Tabla 6: Resultados de la verificación cruzada - Paso 3

Los ejemplos de descripciones donde se encontró el patrón y que coinciden con CustomerID reales incluyen: “sold as 17003?”, “SET 10 CARDS PERFECT POST 17090”, “SET 10 CARD CHRISTMAS WELCOME 17112”, “SET 10 CARDS XMAS CHOIR 17068”, entre otros. Estos casos sugieren que algunos productos en el catálogo contienen números de 5 dígitos como parte de su código o descripción estándar, lo que genera falsos positivos en la detección de concatenación.

PASO 4: Análisis de Control Para validar que el patrón encontrado es específico de la concatenación errónea (y no simplemente códigos de producto), se verificó si también aparece en filas con CustomerID válido.

Interpretación del control:

- **Si el patrón NO aparece en filas válidas:** Evidencia fuerte a favor de H_1
- **Si el patrón SÍ aparece frecuentemente en válidas:** Los 5 dígitos probablemente son códigos de producto, evidencia a favor de H_0

Resultados del grupo control:

Métrica	Valor
Filas válidas con patrón de 5 dígitos	2
Porcentaje de filas válidas con patrón	0.0003 %
Filas nulas con patrón	240
Porcentaje de filas nulas con patrón	0.10 %

Tabla 7: Comparación entre grupo de estudio y grupo control - Paso 4

Ambos ejemplos del grupo control corresponden al mismo producto: “SET 10 CARDS HANGING BAUBLES 17080”. Esto confirma que los números de 5 dígitos encontrados son parte de códigos de producto estándar del catálogo (como la serie 17xxx de tarjetas), y no producto de concatenación accidental.

PASO 5: Visualización y Análisis Estadístico Se generaron visualizaciones para evaluar la magnitud del problema y su relevancia estadística.

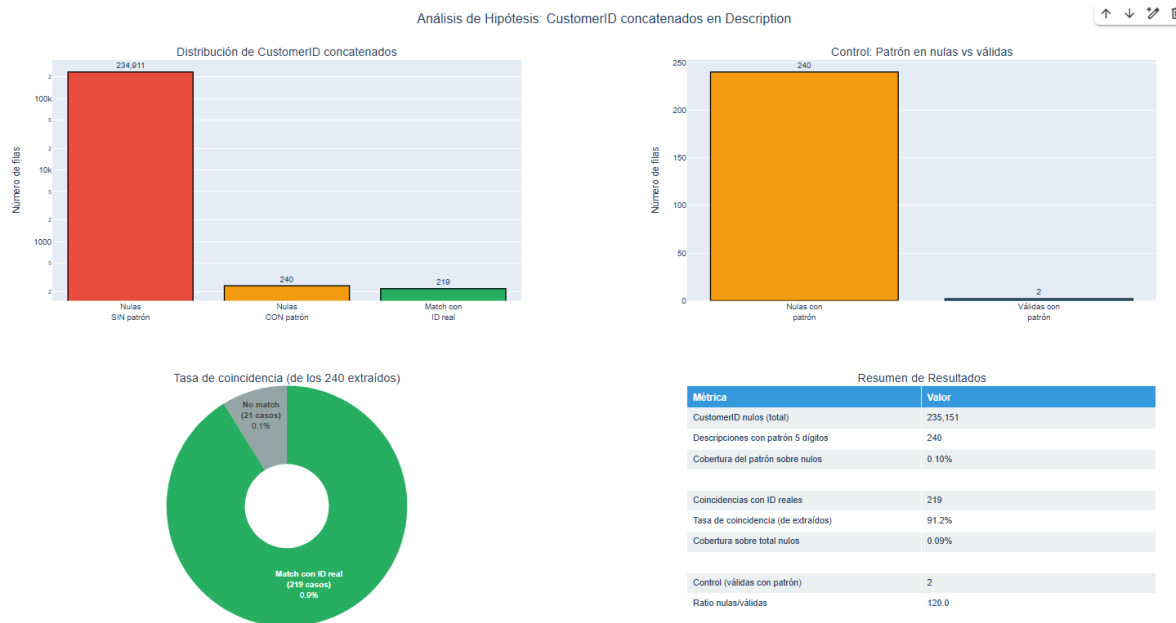


Figura 7: Visualización de resultados del análisis de concatenación de CustomerID

Conclusión

Métrica Clave	Valor
Total CustomerID nulos	235,151
Descripciones con patrón 5 dígitos	240 (0.10 %)
Coincidencias recuperables	219 (0.09 %)
CustomerID no explicados	234,932 (99.91 %)

Tabla 8: Resumen de métricas finales - Hipótesis 1

Decisión: SE RECHAZA H_1 Y SE ACEPTA H_0

Justificación:

Aunque el 91.25 % de los números extraídos coinciden con CustomerID reales, esto representa SOLO el 0.09 % del total de CustomerID nulos (219 de 235,151). La concatenación accidental NO explica la causa principal de los valores nulos.

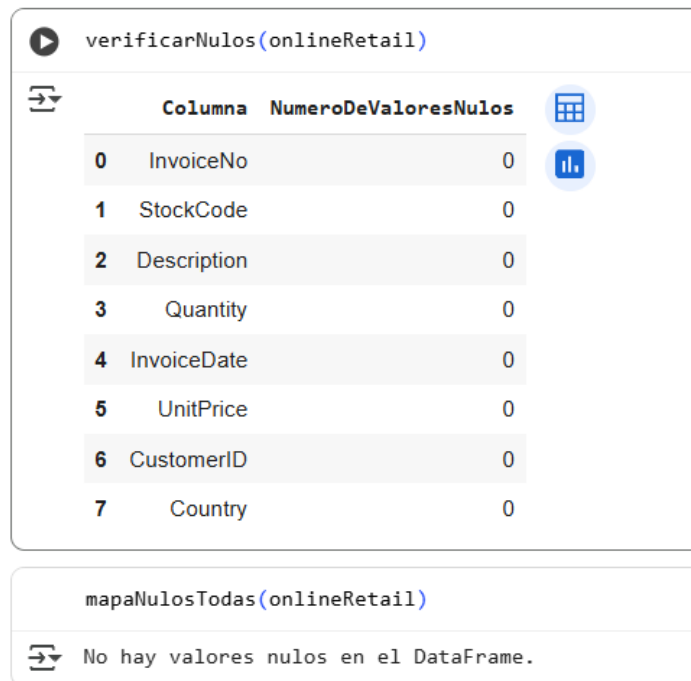
Interpretación estadística:

- **Precisión del patrón:** 91.25 % (de los 240 números extraídos)
- **Cobertura sobre nulos:** 0.09 % (del total de 235,151 nulos)
- **Poder explicativo:** Insuficiente (<1 % threshold)

Adicionalmente, el análisis de control reveló que los números de 5 dígitos encontrados (como la serie 17xxx) son códigos de producto estándar del catálogo, no CustomerID concatenados. La evidencia NO respalda la hipótesis de concatenación masiva. Los valores nulos probablemente se deben a otras causas como: compras sin registro de cliente, ventas B2B sin CustomerID asignado, o errores sistemáticos en el sistema de captura de datos.

3.5. Acciones frente a los datos nulos

Dado que la cantidad de nulos en `CustomerID` es considerable, se decidió eliminarlos mediante la función `dropna()`. Con esto, se asegura que las filas restantes mantengan información íntegra y útil para análisis posteriores, especialmente en tareas de segmentación y estudios de fidelización de clientes. Para la columna `Description`, dado que los nulos son pocos (**1,454** registros), se opta por eliminar esas filas.



The screenshot shows a Jupyter Notebook interface. At the top, there is a play button icon followed by the code `verificarNulos(onlineRetail)`. Below this, a table displays the results of the function. The table has two columns: 'Columna' and 'NumeroDeValoresNulos'. It lists eight columns from the dataset, all of which have a value of 0 in the 'NumeroDeValoresNulos' column, indicating no null values. To the right of the table are two icons: a grid icon and a bar chart icon. Below the table, there is another code cell with the code `mapaNulosTodas(onlineRetail)`. At the bottom, a message box with a double arrow icon states 'No hay valores nulos en el DataFrame.'

	Columna	NumeroDeValoresNulos
0	InvoiceNo	0
1	StockCode	0
2	Description	0
3	Quantity	0
4	InvoiceDate	0
5	UnitPrice	0
6	CustomerID	0
7	Country	0

`mapaNulosTodas(onlineRetail)`

No hay valores nulos en el DataFrame.

Figura 8: Salida de verificación indicando que no hay valores nulos en el dataset *Online Retail II* tras la limpieza.

3.6. ¿Todos los datos están en su formato adecuado?

Durante la revisión de los campos del dataset *Online Retail II*, se observaron las siguientes particularidades:

- **InvoiceNo**: en ciertos casos contiene letras, como la “C”, que identifica facturas canceladas.

Como se muestra en la Figura 9, se puede visualizar la proporción de facturas canceladas frente a las que no lo están.



Figura 9: Número de facturas canceladas (“C”) frente a las facturas no canceladas en el dataset *Online Retail II*.

La Figura 10 muestra la salida de facturas limpias, evidenciando que todas las filas correspondientes a facturas canceladas (con **InvoiceNo** que comienza con “C”) han sido eliminadas del dataset. En total, se removieron alrededor de **34,335 registros**, lo que asegura que el análisis posterior se realice únicamente sobre transacciones válidas y completas, manteniendo la coherencia de los datos y evitando que devoluciones o cancelaciones distorsionen los resultados.

```
No hay facturas canceladas (InvoiceNo comenzando con 'C').
```

Figura 10: Salida de facturas canceladas limpias. *Online Retail II*.

- **UnitPrice**: presenta valores atípicos asociados únicamente a precios nulos (0), los cuales corresponden a devoluciones de productos.

La Figura 11 muestra la distribución de las cantidades, donde se evidencian dichos valores negativos y atípicos.

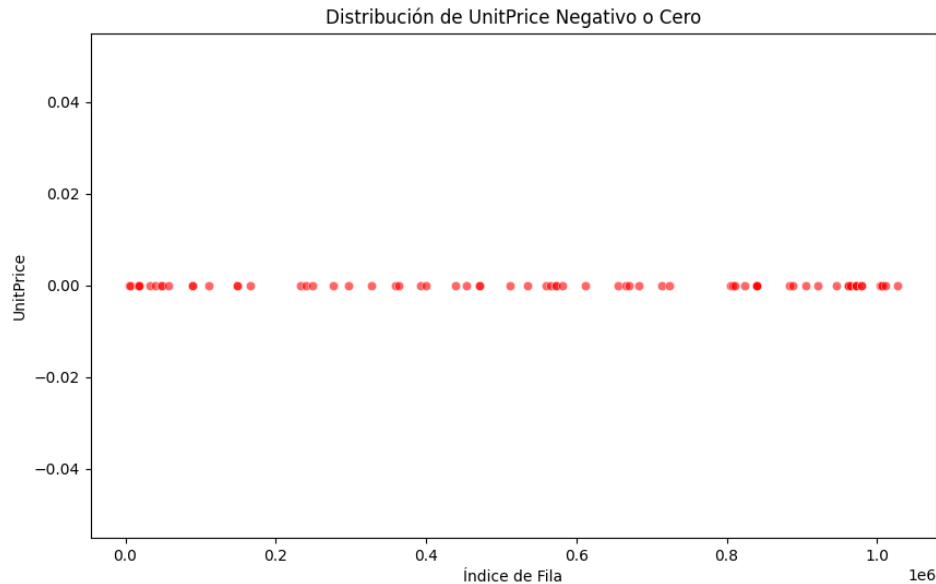


Figura 11: Distribución de la cantidad precio unitario inválidos en el dataset *Online Retail II*.

Para garantizar la consistencia de los análisis posteriores, se aplicó un filtro que conserva únicamente las transacciones con valores válidos:

- Se eliminaron los registros con **UnitPrice** menores o iguales a cero.

Este filtrado se implementó mediante un **query** que asegura que todas las filas restantes tengan cantidades y precios positivos. De esta forma, se mantiene la coherencia de los datos y se evita que valores atípicos o devoluciones afecten los resultados del análisis.

La Figura 12 muestra la nueva distribución de las cantidades y precios unitarios después del filtrado.

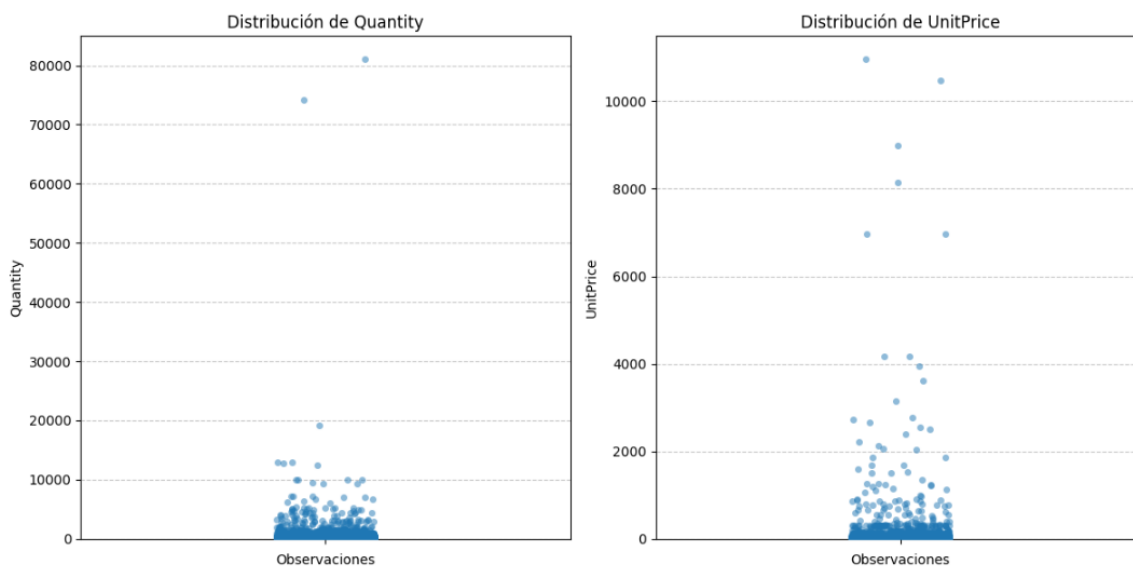


Figura 12: Distribución de la cantidad de productos por transacción y del precio unitario después del filtrado en el dataset *Online Retail II*.

Adicionalmente, en el atributo **Country** se detectaron valores con codificación no estándar, como se muestra en la Tabla 9.

País	Frecuencia	Observación
EIRE	15,565	Corresponde a Irlanda, requiere normalización a "Ireland".
Channel Islands	1,551	Región dependiente del Reino Unido, se puede agrupar bajo "United Kingdom".
Unspecified	518	Registros sin país especificado, se sugiere eliminar o agrupar en "Others".
USA	409	Corresponde a Estados Unidos, requiere uniformización a "United States of America".
RSA	122	Corresponde a South Africa (República de Sudáfrica), debe normalizarse.
European Community	60	Valor genérico sin país específico, no aporta información clara; se sugiere eliminar.
West Indies	54	Denominación ambigua que agrupa varias islas del Caribe; difícil de normalizar.
Korea	53	Ambiguo: no diferencia entre Corea del Sur y Corea del Norte; requiere revisión manual.
Czech Republic	25	Corresponde a "Czechia", debe normalizarse.

Tabla 9: Países con codificación inusual, ambigua o baja frecuencia en el atributo **Country**.

A pesar de tratarse de países poco frecuentes o con codificación atípica, solo se eliminarán las filas correspondientes a los registros agrupados como “Others”.

Para garantizar la coherencia de los datos en el atributo **Country**, se aplicó un proceso de **normalización de países**. Este procedimiento consistió en reemplazar los valores poco usuales o no estandarizados por su equivalente correcto.

De esta manera, se asegura una mayor consistencia en los valores categóricos y se evitan problemas posteriores en el análisis derivados de nombres duplicados o poco claros, como lo podemos observar en la figura 13.

```
def normalizarPaíses(dataFrame, columnaPaís):
    mapeoPaíses = {
        'EIRE': 'Ireland',
        'Channel Islands': 'United Kingdom',
        'Unspecified': "Others",
        'USA': 'United States of America',
        'European Community': "Others",
        'West Indies': "Others",
        'RSA': 'South Africa',
        'Czech Republic': 'Czechia',
        'Korea': 'South Korea'
    }
    dataFrame[columnaPaís] = dataFrame[columnaPaís].replace(mapeoPaíses)
```

Ejecutamos

```
normalizarPaíses(onlineRetail, 'Country')
```

Eliminamos los registros "Others", dado que no tienen un país de referencia.

```
onlineRetail.drop(onlineRetail[onlineRetail['Country'] == 'Others'].index, inplace=True)
```

Verificamos:

```
PaísesNoReconocidos = AnalizarPaíses(onlineRetail, world)
if not PaísesNoReconocidos:
    print("No se encontraron países no reconocidos.")
else:
    for país, contador in PaísesNoReconocidos.items():
        print(f'{país}: {contador}')
```

No se encontraron países no reconocidos.

Figura 13: Proceso de normalización en Country.

3.7. Análisis de outliers

3.7.1. ¿Cuáles son los Outliers?

Para identificar valores atípicos en el dataset *Online Retail*, se analizaron principalmente dos variables: la cantidad de productos (**Quantity**) y el precio unitario (**UnitPrice**).

Estos valores extremos pueden deberse a errores de registro, promociones excepcionales, pedidos grandes de clientes mayoristas o devoluciones de productos.

La Figura 14 muestra la distribución de **Quantity**, donde cada punto representa la cantidad de un producto en una transacción. Los puntos rojos representan las observaciones que se encuentran fuera del rango intercuartílico, indicando pedidos inusualmente grandes que podrían corresponder a errores de ingreso de datos o compras mayoristas.

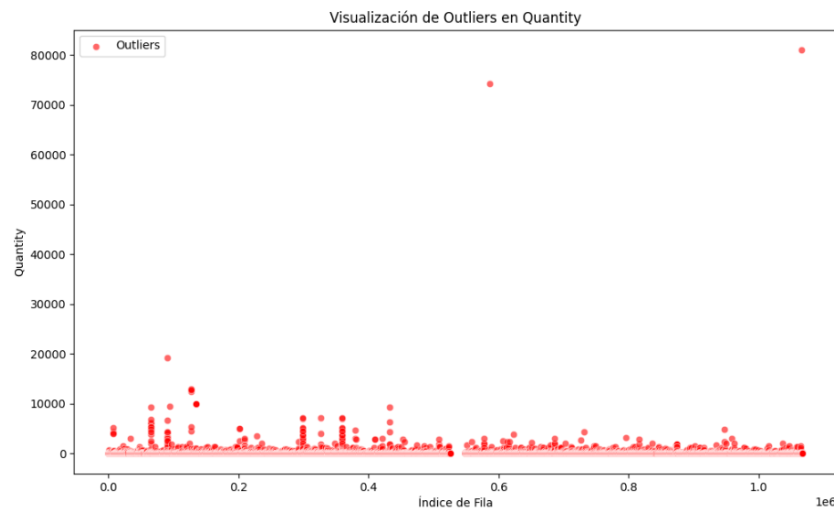


Figura 14: Visualización de outliers en **Quantity**. Los puntos rojos representan valores atípicos identificados mediante el rango intercuartílico.

De manera similar, la Figura 15 muestra los outliers de **UnitPrice**. Cada punto corresponde al precio unitario de un producto en una transacción. Los puntos rojos representan precios excepcionalmente bajos o altos que se encuentran fuera del rango típico.

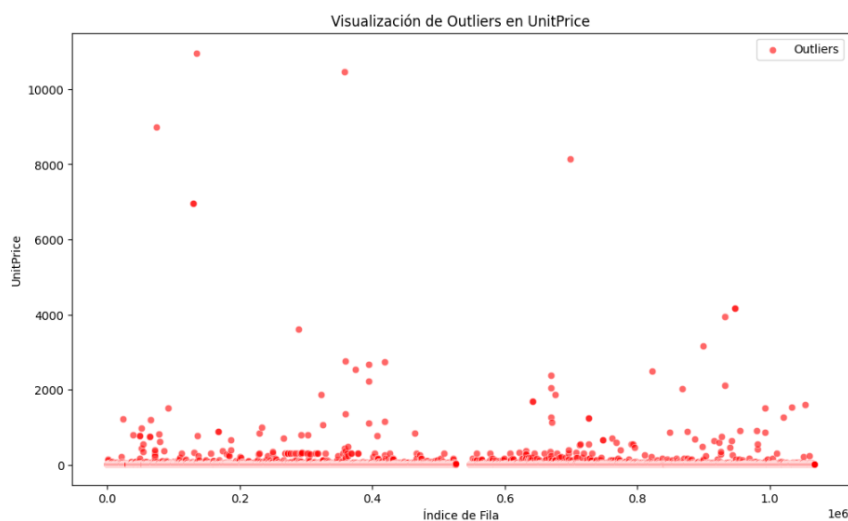


Figura 15: Visualización de outliers en **UnitPrice**. Los puntos rojos representan valores atípicos identificados mediante el rango intercuartílico.

Como referencia, se presentan también los gráficos normales sin resaltar outliers. La Figura 16 muestra la distribución completa de **Quantity**, evidenciando que la mayoría de las transacciones se concentran en valores bajos y que los outliers detectados previamente son claramente atípicos.

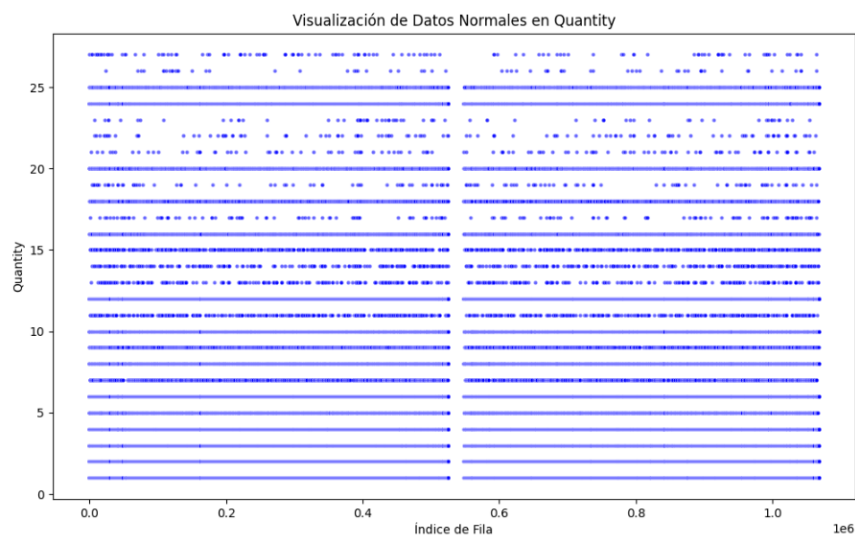


Figura 16: Distribución general de **Quantity** sin resaltar outliers. Se observa la concentración principal de transacciones en cantidades bajas.

De igual forma, la Figura 17 muestra la distribución general de `UnitPrice`. Aquí se puede apreciar la mayoría de precios unitarios dentro de un rango esperado, mientras que los outliers previamente identificados sobresalen como valores extremos fuera de la tendencia general.

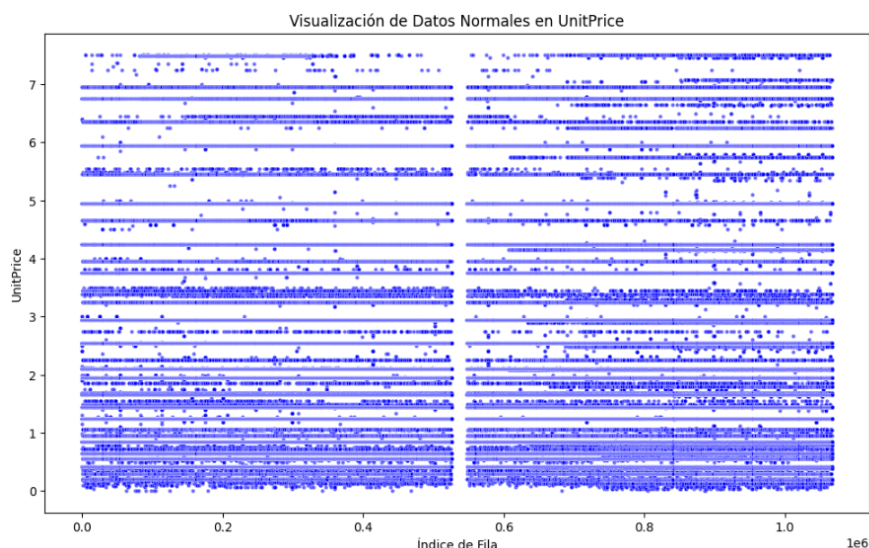


Figura 17: Distribución general de `UnitPrice` sin resaltar outliers. La mayoría de precios se encuentran dentro de un rango típico, destacando los valores extremos previamente identificados.

3.7.2. ¿Podemos eliminarlos? ¿Es importante conservarlos?

Como señala Khandelwal [5], en el contexto del *retail online* no resulta conveniente aplicar de manera automática funciones de limpieza como `RemoveOutliers` que eliminan o normalizan todos los valores extremos. A diferencia de otras disciplinas, donde los outliers suelen considerarse ruido estadístico, en el comercio electrónico muchos de estos valores representan información de alto valor estratégico para la empresa.

Por ejemplo, cantidades excepcionalmente grandes en una transacción pueden corresponder a clientes mayoristas o corporativos, mientras que precios unitarios elevados pueden estar asociados a productos premium o de lujo. Estos casos no deben considerarse simples anomalías, sino indicadores de segmentos de clientes VIP o de oportunidades de negocio relevantes. En este sentido, la visión propuesta por Khandelwal [5] resalta la importancia de interpretar los datos desde una perspectiva comercial antes de decidir cualquier proceso de normalización o eliminación de outliers.

Por tanto, más que aplicar de manera mecánica un procedimiento de limpieza que descarte estos registros, resulta fundamental analizarlos en función de su significado de negocio. Conservar este tipo de información permite identificar patrones de consumo diferenciados, segmentar clientes de alto valor y diseñar estrategias de marketing más precisas.

4. Exploración

4.1. ¿Cuántos productos, clientes y países existen?

El conjunto de datos *Online Retail II* contiene información detallada sobre las transacciones de una tienda en línea. En la **Tabla 10** se presenta un resumen general con la cantidad total de productos únicos, clientes distintos y países registrados, lo que evidencia el alcance global y la diversidad del negocio analizado.

Categoría	Cantidad
Productos únicos	5283
Clientes distintos	5870
Países registrados	37

Tabla 10: Resumen general del dataset *Online Retail II*

4.2. ¿Cuáles son los productos más vendidos?

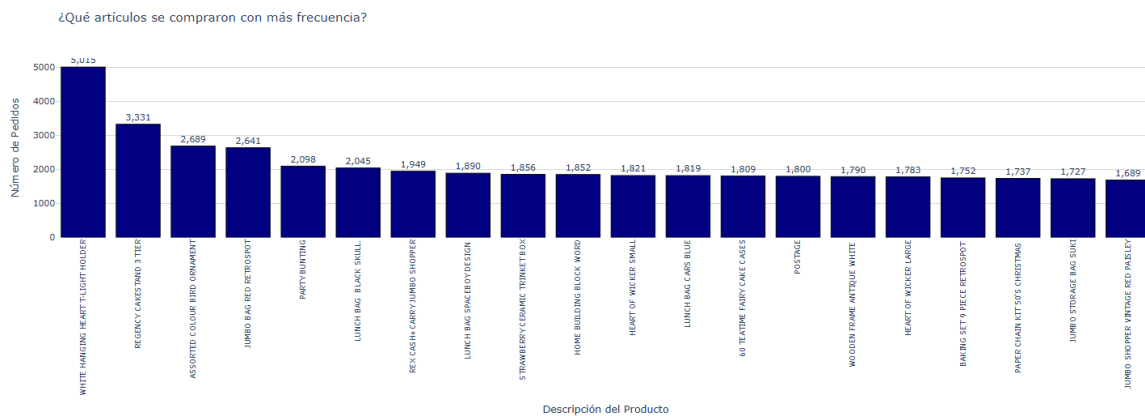


Figura 18: Top 20 productos más pedidos en términos de número de transacciones.

La Figura 18 muestra los productos más pedidos según el número de transacciones registradas en el dataset *Online Retail II*. Se observa que el artículo **WHITE HANGING HEART T-LIGHT HOLDER** es el más popular, con un total de **5,015 pedidos**, seguido por otros productos que concentran un menor número de transacciones. Esto refleja un patrón típico en comercio minorista donde pocos productos concentran la mayoría de las compras.

Conclusión: los productos más frecuentemente pedidos no siempre coinciden con los de mayor volumen en cantidad. La frecuencia de compra permite identificar los artículos de alta demanda recurrente, lo cual es clave para estrategias de reposición y marketing.

4.3. ¿Cuáles son los productos más vendidos por cantidad?



Figura 19: Top 20 productos más vendidos por cantidad total comprada.

La Figura 19 presenta los productos que se han vendido en mayor cantidad acumulada. El artículo **WORLD WAR 2 GLIDERS ASSTD DESINGS** encabeza la lista, evidenciando que aunque algunos productos puedan no ser los más frecuentemente pedidos, las cantidades adquiridas por pedido pueden ser significativamente mayores. El resto de productos del top 20 muestra cantidades relevantes, indicando que la venta en volumen también se concentra en unos pocos artículos.

Observación: Comparando con la subsección anterior, se nota que los productos más vendidos por frecuencia (número de pedidos) no siempre coinciden con los productos más vendidos por cantidad (volumen total de unidades). Esto ocurre porque un producto puede ser muy popular en número de pedidos pero vender pocas unidades por transacción, mientras que otro producto puede ser comprado en menor número de pedidos pero con grandes cantidades por pedido. Esta distinción es relevante para la gestión del inventario y la planificación de stock, ya que permite identificar tanto los artículos de alta demanda recurrente como aquellos que generan mayor volumen de ventas.

4.4. ¿Siguen alguna distribución?

Para analizar la distribución de los datos en el dataset *Online Retail II*, se generó un resumen estadístico de las principales variables numéricas, cuyos resultados se presentan en la Tabla 11.

Métrica	Quantity	UnitPrice
count	778,793	778,793
mean	13.49	3.22
std	145.91	29.69
min	1.00	0.001
25 %	2.00	1.25
50 %	6.00	1.95
75 %	12.00	3.75
max	80,995.00	10,953.50

Tabla 11: Resumen estadístico actualizado de las variables **Quantity** y **UnitPrice** en el dataset *Online Retail II*.

A continuación se presentan interpretaciones obtenidas exclusivamente a partir de los valores de la Tabla 11.

■ **Quantity:**

- La media es 13.49 y la mediana 6; como $\text{media} > \text{mediana}$, la distribución presenta **sesgo positivo** (cola a la derecha).
- La desviación estándar es 145.91, aproximadamente $\frac{145,91}{13,49} \approx 10,8$ veces la media; esto indica una **alta dispersión** respecto a los valores centrales.
- El máximo (80,995) es extremadamente mayor que los cuantiles: $\frac{80\,995}{6} \approx 13,499$ veces la mediana y $\frac{80\,995}{12} \approx 6,750$ veces el percentil 75. Esto confirma la existencia de **outliers extremos** que elevan la media y la varianza.
- Conclusión: **Quantity** no sigue una distribución normal; su patrón es característico de datos transaccionales con muchas compras pequeñas y pocos pedidos masivos.

■ **UnitPrice:**

- La mediana es 1.95 y la media 3.22; nuevamente $\text{media} > \text{mediana}$, indicando **sesgo positivo**.
- La desviación estándar es 29.69, que equivale a $\frac{29,69}{3,22} \approx 9,2$ veces la media; los valores atípicos dominan la varianza.
- El máximo (10,953.50) es desproporcionado respecto a los cuantiles: $\frac{10\,953,50}{1,95} \approx 5,617$ veces la mediana y $\frac{10\,953,50}{3,75} \approx 2,921$ veces el percentil 75. Esto señala precios fuera de rango usual, posiblemente por errores de captura o artículos muy especiales.
- Conclusión: **UnitPrice** muestra una distribución fuertemente asimétrica a la derecha; el uso de la mediana o transformaciones como logaritmos es más adecuado para describir su comportamiento.

4.5. ¿Entre qué rangos están los datos?

En esta sección se analizan los valores mínimos y máximos de las variables del dataset **Online Retail II**, con el fin de identificar posibles inconsistencias o rangos relevantes para el análisis.

- **InvoiceNo:** no resulta de interés calcular valores mínimo y máximo, ya que se trata de identificadores de facturas.
- **StockCode:** al ser códigos de productos, tampoco es relevante observar su mínimo y máximo.
- **CustomerID:** no resulta de interés calcular valores mínimo y máximo, ya que se trata de identificadores de clientes.
- **Description:** corresponde a los nombres de los productos, por lo que no aplica un análisis de rangos.
- **Quantity:** se observa el rango de cantidades de productos por transacción.

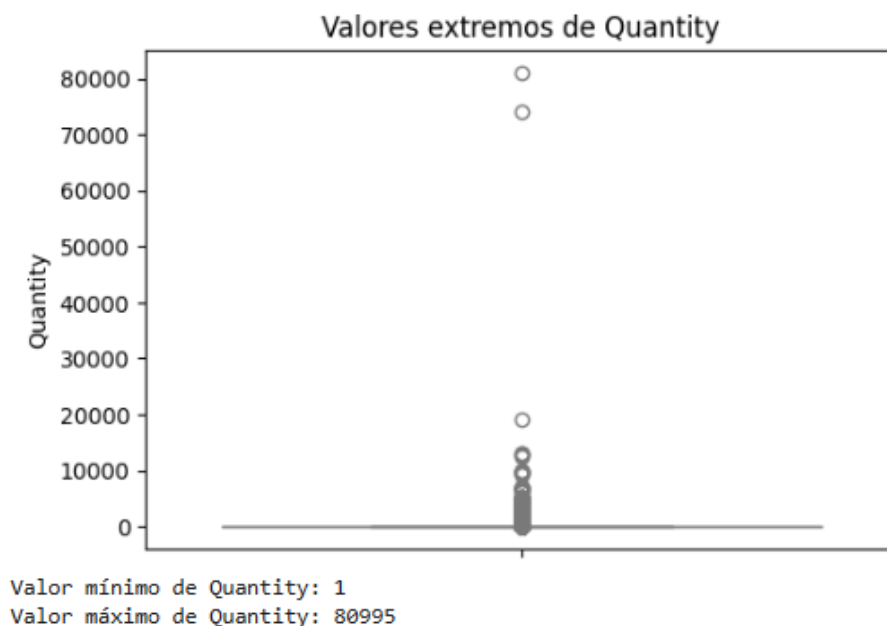


Figura 20: Distribución de la variable Quantity.

Este valor máximo resulta muy elevado y puede corresponder a un caso atípico o a un error de registro, ya que es inusual que una sola transacción involucre más de 80 mil unidades. El valor mínimo de 1 es coherente, pues representa la venta de un solo producto.

Top 10 productos más vendidos (por cantidad).

A continuación, se muestran los diez productos con mayor volumen total de ventas. Se observa que el artículo *PAPER CRAFT, LITTLE BIRDIE* lidera ampliamente con más de 80 mil unidades vendidas.

Descripción	Precio Unitario (£)	Cantidad Total
PAPER CRAFT, LITTLE BIRDIE	2.08	80,995
MEDIUM CERAMIC TOP STORAGE JAR	1.04	76,087
WHITE HANGING HEART T-LIGHT HOLDER	2.55	57,577
WORLD WAR 2 GLIDERS ASSTD DESIGNS	0.21	45,660
ASSORTED COLOUR BIRD ORNAMENT	1.69	44,527
WORLD WAR 2 GLIDERS ASSTD DESIGNS	0.29	34,078
ASSORTED COLOUR BIRD ORNAMENT	1.45	32,581
SMALL POPCORN HOLDER	0.72	30,814
MINI PAINT SET VINTAGE	0.65	30,627
60 TEATIME FAIRY CAKE CASES	0.55	28,387

Tabla 12: Top 10 productos más vendidos por cantidad en el dataset *Online Retail II*.

- UnitPrice: rango de precios unitarios de los productos.

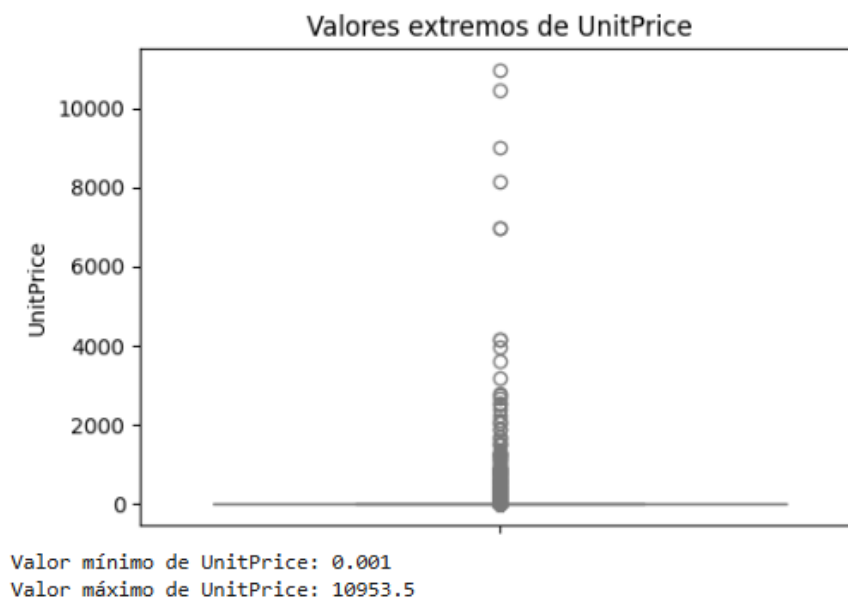


Figura 21: Distribución de la variable UnitPrice.

El valor mínimo sugiere posibles registros incorrectos o precios mal capturados (un producto no puede costar 0.001). El máximo también es inusualmente alto, lo cual podría reflejar un error de carga o un producto premium extremadamente costoso.

El análisis de los precios unitarios extremos permite identificar posibles valores atípicos dentro del conjunto de datos. El valor mínimo sugiere posibles registros incorrectos o precios mal capturados (un producto no puede costar £0.001). El máximo también es inusualmente alto, lo cual podría reflejar un error de carga o un producto premium extremadamente costoso.

Descripción	UnitPrice (£)	Cantidad
Bank Charges	0.001	1
PADS TO MATCH ALL CUSHIONS	0.001	17

Tabla 13: Productos con los precios unitarios mínimos en el dataset *Online Retail II*.

Descripción	UnitPrice (£)	Cantidad
Manual	10,953.50	1

Tabla 14: Productos con los precios unitarios máximos en el dataset *Online Retail II*.

Estos resultados evidencian que existen valores extremos que podrían corresponder a errores de registro o situaciones excepcionales dentro del conjunto de datos, por lo que deben considerarse cuidadosamente durante la limpieza y el análisis posterior.

- InvoiceDate: rango temporal de las transacciones en el dataset.

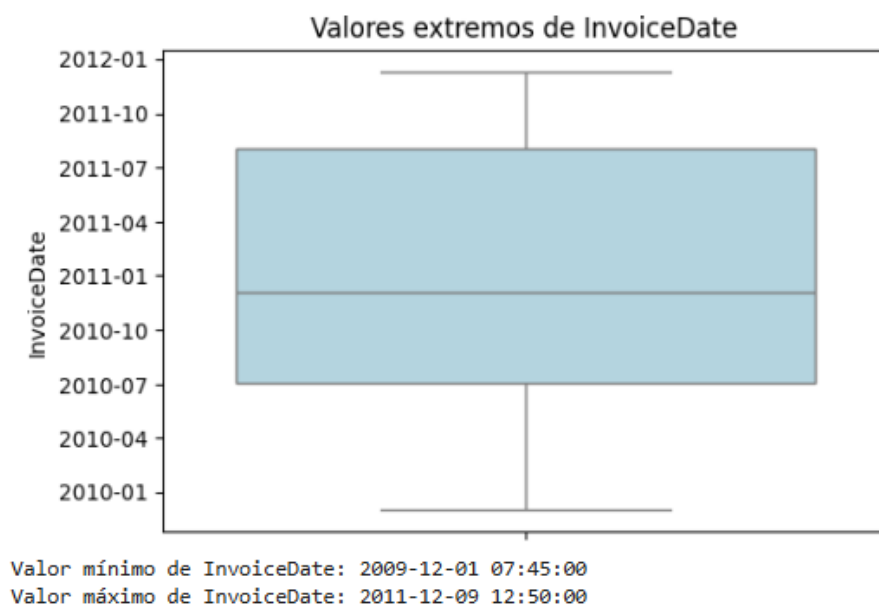


Figura 22: Distribución temporal y valores extremos de InvoiceDate.

Este rango confirma que los datos abarcan un año completo de operaciones. No se detectan valores anómalos en este caso.

- **Country:** el dataset registra transacciones de **37 países**. En la Figura 23 se presenta un mapa donde se resaltan los países incluidos. El predominio de transacciones en Reino Unido sugiere que la empresa tiene allí su principal mercado, mientras que las ventas internacionales son considerablemente menores y están más dispersas geográficamente.

Asignación de números a países (orden alfabético):

- | | | |
|--------------|-----------------|------------------------------|
| 1. Australia | 14. Iceland | 27. Saudi Arabia |
| 2. Austria | 15. Ireland | 28. Singapore |
| 3. Bahrain | 16. Israel | 29. South Africa |
| 4. Belgium | 17. Italy | 30. South Korea |
| 5. Brazil | 18. Japan | 31. Spain |
| 6. Canada | 19. Lebanon | 32. Sweden |
| 7. Cyprus | 20. Lithuania | 33. Switzerland |
| 8. Czechia | 21. Malta | 34. Thailand |
| 9. Denmark | 22. Netherlands | 35. United Arab Emirates |
| 10. Finland | 23. Nigeria | 36. United Kingdom |
| 11. France | 24. Norway | 37. United States of America |
| 12. Germany | 25. Poland | |
| 13. Greece | 26. Portugal | |

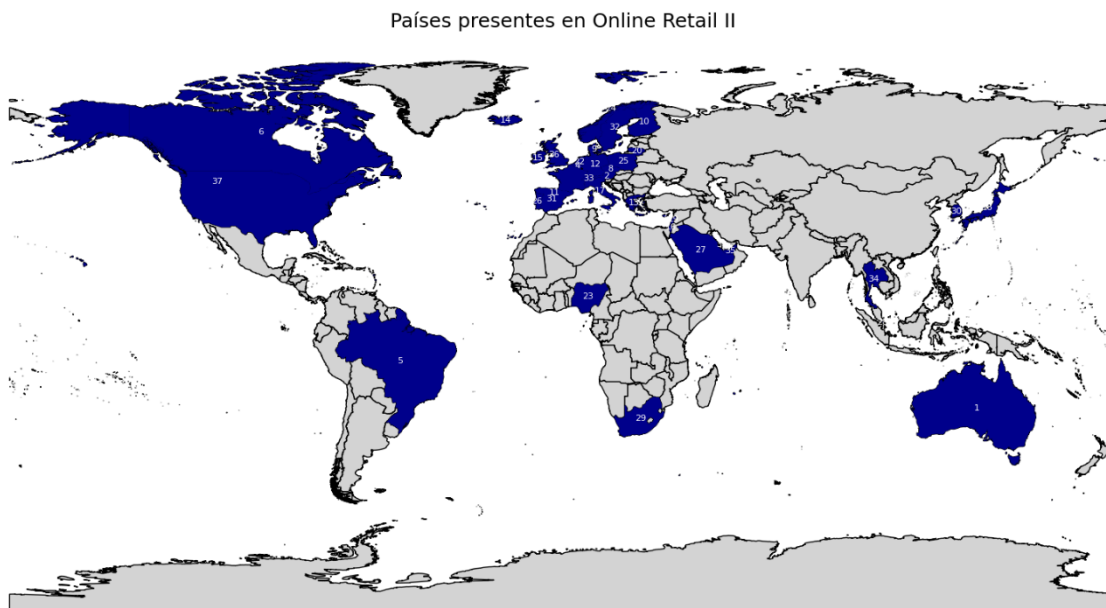


Figura 23: Mapa de países con transacciones registradas en el dataset *Online Retail II*.

4.6. ¿Cuál es la tendencia diaria de ventas?

Para analizar el comportamiento temporal de las ventas, se generaron visualizaciones interactivas a partir de la variable `InvoiceDate`. Estas permiten observar la evolución de las ventas **diarias** durante todo el periodo disponible en el dataset *Online Retail II*, comprendido entre los años 2009 y 2011.

En las visualizaciones se puede filtrar fácilmente por año, lo que facilita identificar patrones estacionales, fluctuaciones diarias y picos de ventas específicos. A continuación, se presentan las figuras correspondientes:

- **Figura 24:** Tendencia general de ventas diarias considerando todos los años del periodo analizado.
- **Figura 25:** Evolución de las ventas diarias durante el año 2009.
- **Figura 26:** Evolución de las ventas diarias durante el año 2010.
- **Figura 27:** Evolución de las ventas diarias durante el año 2011.

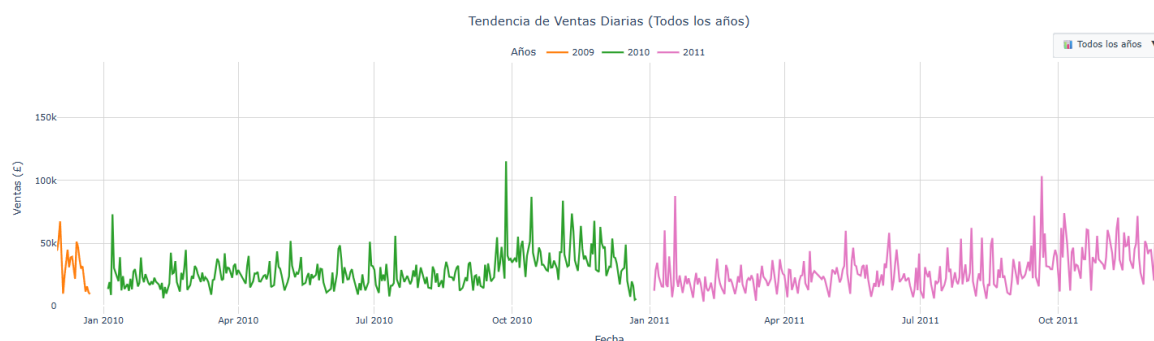


Figura 24: Tendencia general de ventas diarias.

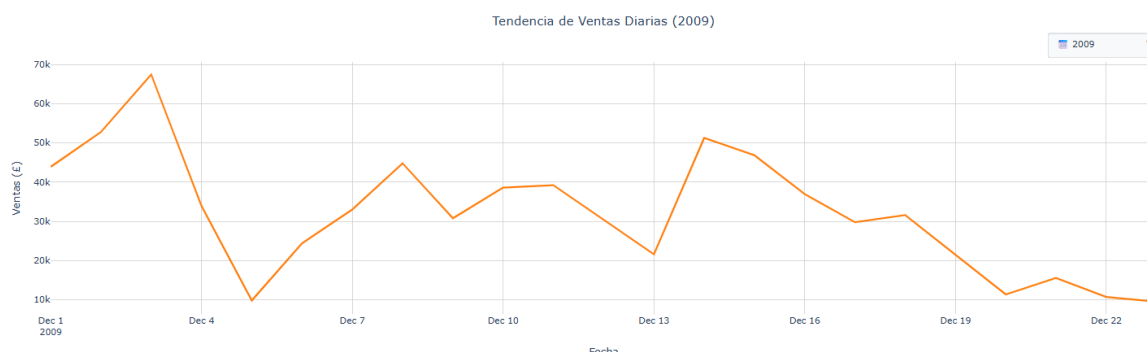


Figura 25: Tendencia de ventas diarias durante el año 2009.

Durante el año 2009 se observa una actividad comercial limitada, con un volumen de ventas relativamente bajo en comparación con los años siguientes. Destaca el 5 de diciembre de 2009, fecha en la que se registraron aproximadamente £9,803 en ventas, siendo uno de los pocos picos notorios de ese año.

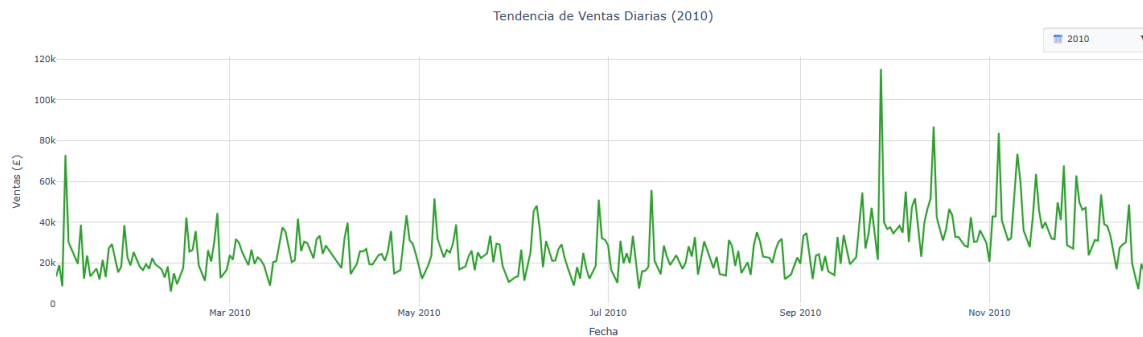


Figura 26: Tendencia de ventas diarias durante el año 2010.

En el año 2010, las ventas presentan un comportamiento más activo, con varios picos a lo largo del año. Se observa un aumento destacable el 7 de enero, y el pico más pronunciado ocurre el 27 de septiembre de 2010, alcanzando más de £115,000 en un solo día. Esto sugiere un incremento en la demanda o la realización de algún evento comercial específico durante ese periodo.

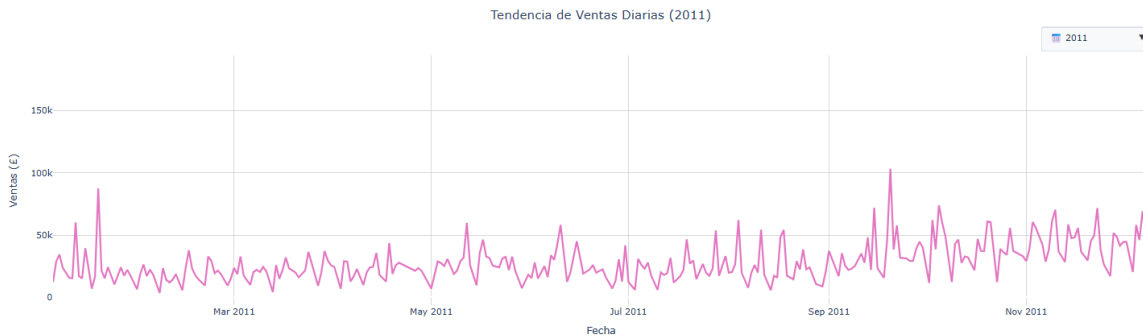


Figura 27: Tendencia de ventas diarias durante el año 2011.

Finalmente, en 2011 se observa una actividad considerablemente mayor, con múltiples picos a lo largo del año. Por ejemplo, el 18 de enero se registraron cerca de £87,000 en ventas, seguido por £13,000 el 13 de mayo y £58,000 el 10 de junio. Asimismo, el 21 de septiembre se alcanzaron alrededor de £21,000, posiblemente vinculado a alguna campaña estacional como la llegada de la primavera. No obstante, el pico más alto de todo el periodo analizado ocurre el 9 de diciembre de 2011, con un volumen de ventas superior a £174,000, reflejando un fuerte impulso comercial hacia el cierre del año.

En general, se aprecia una tendencia de crecimiento progresivo entre 2009 y 2011, con una marcada estacionalidad hacia los últimos meses de cada año, posiblemente relacionada con campañas navideñas y el incremento de la demanda en temporada alta.

4.7. ¿Cuál es la tendencia mensual de ventas?

Para complementar el análisis temporal, se agruparon las ventas a nivel mensual con el fin de observar la evolución del volumen de ingresos a lo largo de los tres años registrados en el dataset *Online Retail II*.

En las visualizaciones se representan las ventas totales por mes y se incluyen filtros interactivos por año, permitiendo comparar fácilmente los patrones de comportamiento y los picos de ventas más relevantes. A continuación, se presentan las figuras generadas:

- **Figura 28:** Tendencia general de ventas mensuales considerando todos los años.
- **Figura 29:** Evolución de las ventas mensuales durante el año 2009.
- **Figura 30:** Evolución de las ventas mensuales durante el año 2010.
- **Figura 31:** Evolución de las ventas mensuales durante el año 2011.

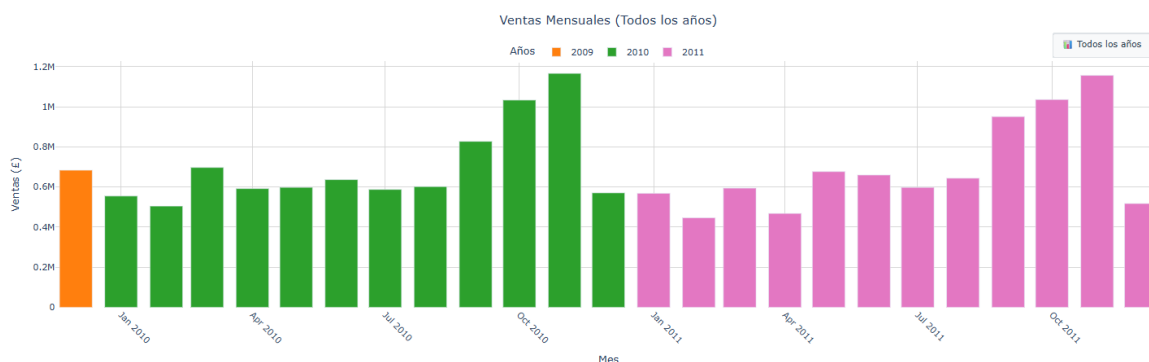


Figura 28: Tendencia general de ventas mensuales.

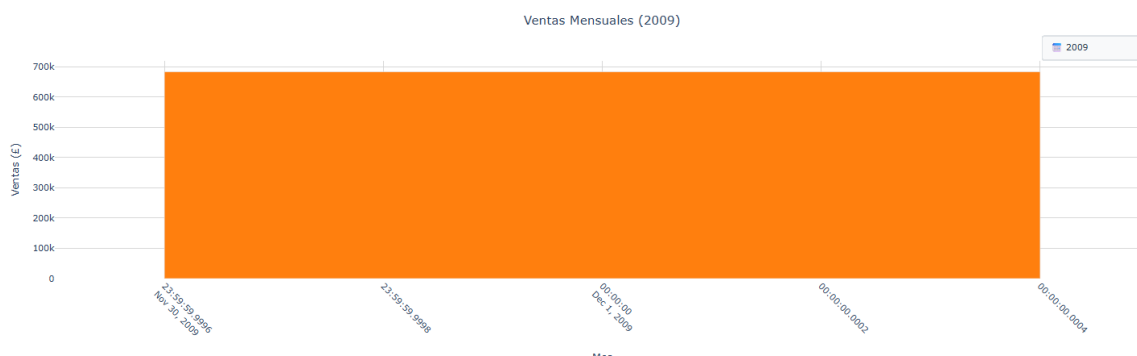


Figura 29: Tendencia de ventas mensuales durante el año 2009.

Durante el año 2009, solo se cuenta con registros de un único mes, en el cual las ventas alcanzaron aproximadamente £683,000. Esto refleja una actividad comercial limitada, probablemente debido a que el conjunto de datos comienza en los últimos meses de dicho año.

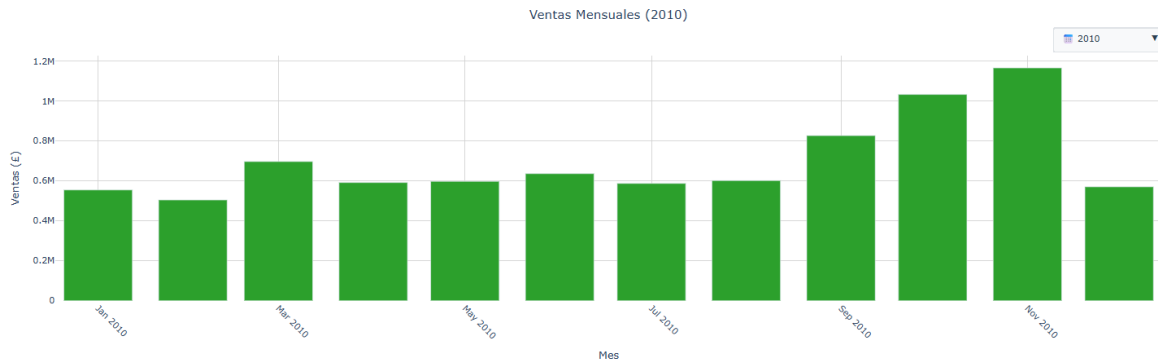


Figura 30: Tendencia de ventas mensuales durante el año 2010.

En el año 2010, las ventas muestran un comportamiento más definido a lo largo de los meses. El valor más bajo se observa en febrero, con un total aproximado de £504,000, mientras que el pico máximo se alcanza en noviembre, con cerca de £1,166,159, evidenciando un incremento significativo en las ventas durante la temporada de fin de año. Esta concentración de ventas en noviembre sugiere la posible influencia de eventos comerciales o campañas prenavideñas.

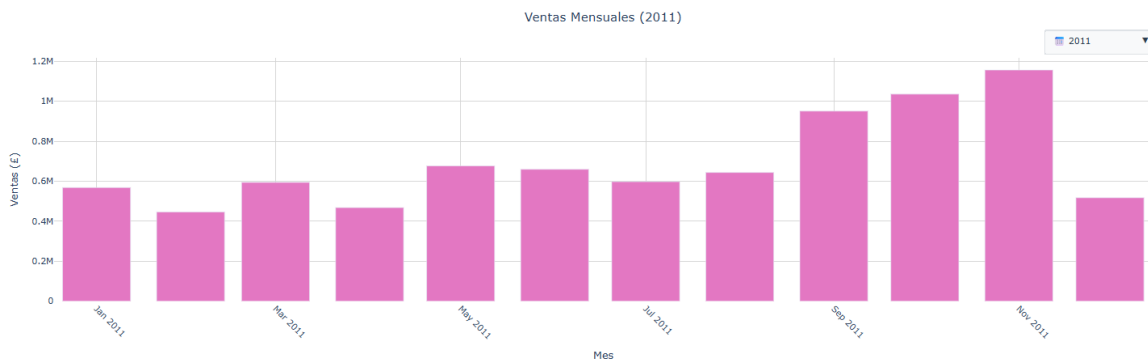


Figura 31: Tendencia de ventas mensuales durante el año 2011.

En el año 2011 se mantiene un patrón similar al del año anterior: el mes de febrero registra el menor volumen de ventas, mientras que noviembre vuelve a ser el mes con el valor máximo, reafirmando la fuerte estacionalidad del negocio en los últimos meses del año. Este comportamiento indica una tendencia recurrente de aumento en las ventas hacia el cierre anual, posiblemente asociada a las festividades de fin de año y a campañas promocionales específicas.

En general, la comparación entre los tres años revela una clara tendencia al crecimiento del volumen total de ventas, junto con una marcada concentración de ingresos durante los meses de noviembre y diciembre, lo que sugiere la existencia de una estacionalidad pronunciada en el comportamiento de compra de los clientes.

4.8. ¿Cuál es el número de transacciones por horas?

Para profundizar en el análisis temporal, se evaluó la distribución del número de transacciones por cada hora del día, considerando tanto el comportamiento agregado de todos los años como el de manera individual para cada uno (2009, 2010 y 2011).

En las siguientes figuras se presentan los mapas de calor (*heatmaps*) que reflejan la intensidad de transacciones por hora y mes:

- **Figura 32:** Número total de transacciones por horas considerando todos los años.
- **Figura 33:** Distribución de transacciones por horas en el año 2009.
- **Figura 34:** Distribución de transacciones por horas en el año 2010.
- **Figura 35:** Distribución de transacciones por horas en el año 2011.

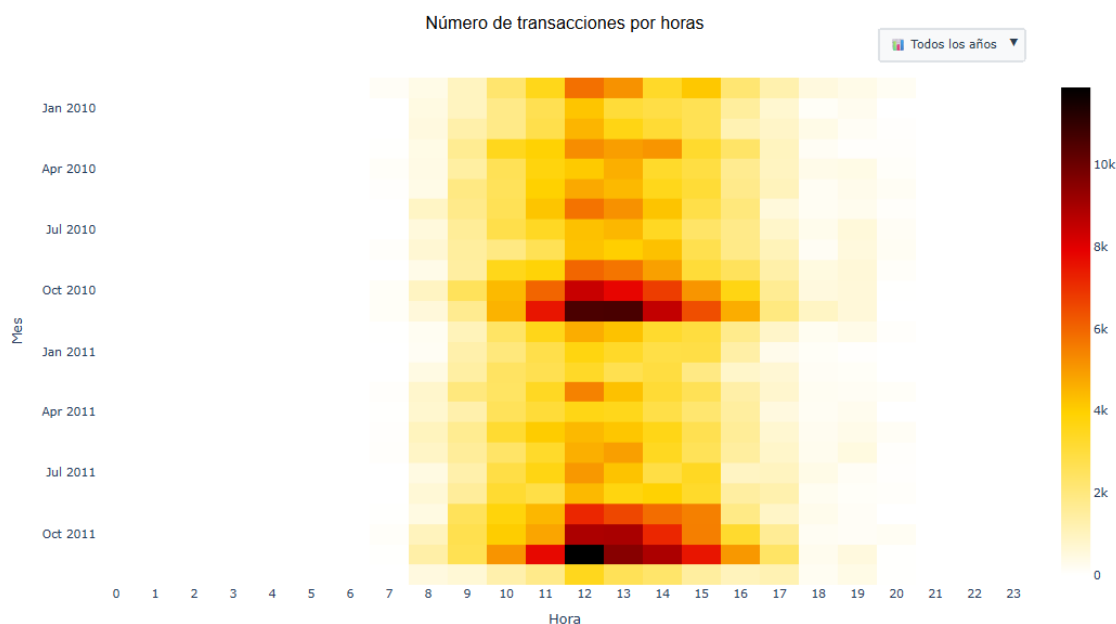


Figura 32: Número de transacciones por horas considerando todos los años.

En la figura general se observa que el mayor número de transacciones se concentra principalmente entre las **12:00 p.m. y las 4:00 p.m.**, con un pico máximo alrededor del **mediodía (12:00 p.m.)**, alcanzando aproximadamente **11,894 transacciones**.

Este comportamiento se mantiene de manera consistente a lo largo de los distintos años, lo que sugiere un patrón de compra recurrente durante el horario de mayor actividad laboral o disponibilidad de los clientes.

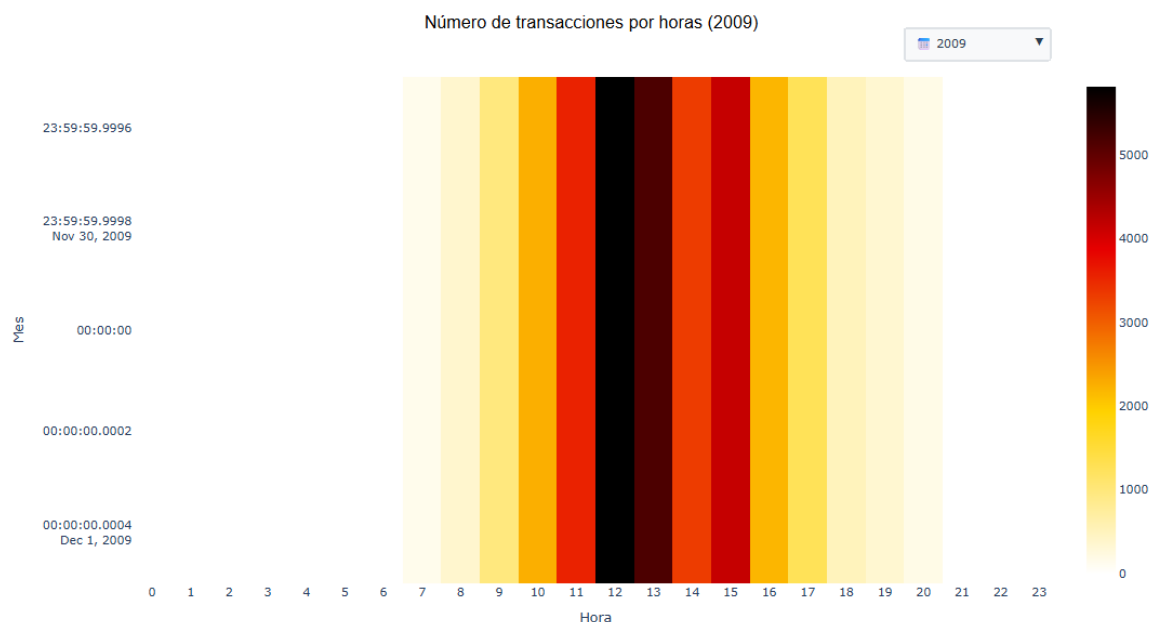


Figura 33: Distribución de transacciones por horas en el año 2009.

Durante el año 2009, aunque el volumen total de registros es menor, se mantiene la tendencia de mayor actividad entre las 12:00 p.m. y las 4:00 p.m., con un pico central alrededor del mediodía.

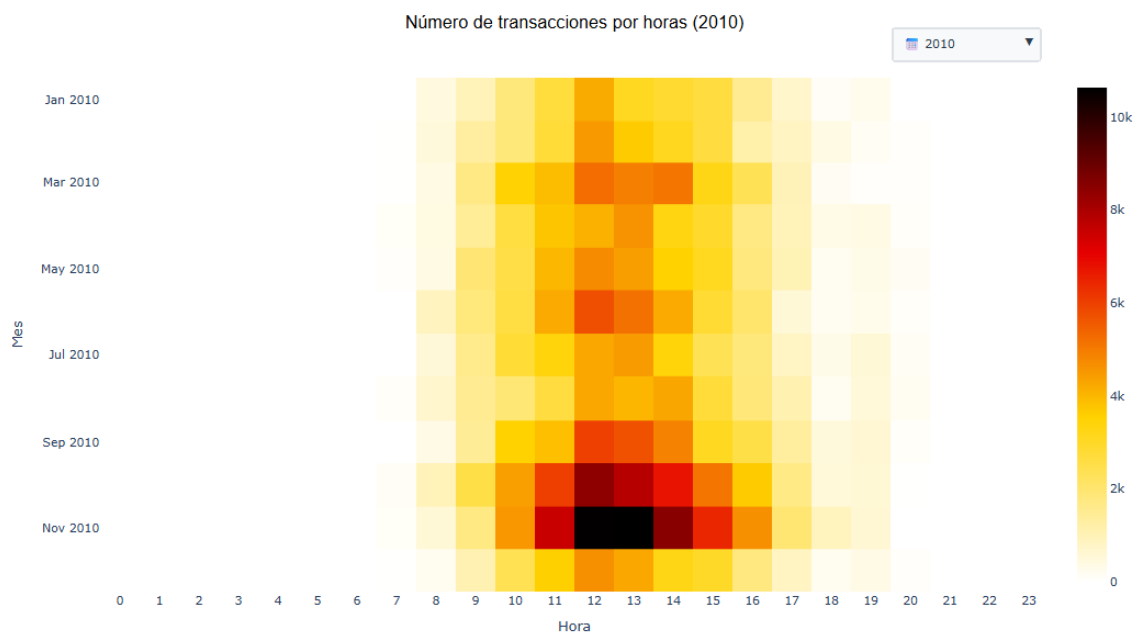


Figura 34: Distribución de transacciones por horas en el año 2010.

En el año 2010 se aprecia un patrón similar: las horas de mayor movimiento comercial se concentran también entre las 12:00 p.m. y las 4:00 p.m., evidenciando que los clientes realizan la mayoría de las

compras durante el horario de mediodía, con un pico marcado cercano a las 12:00 p.m.

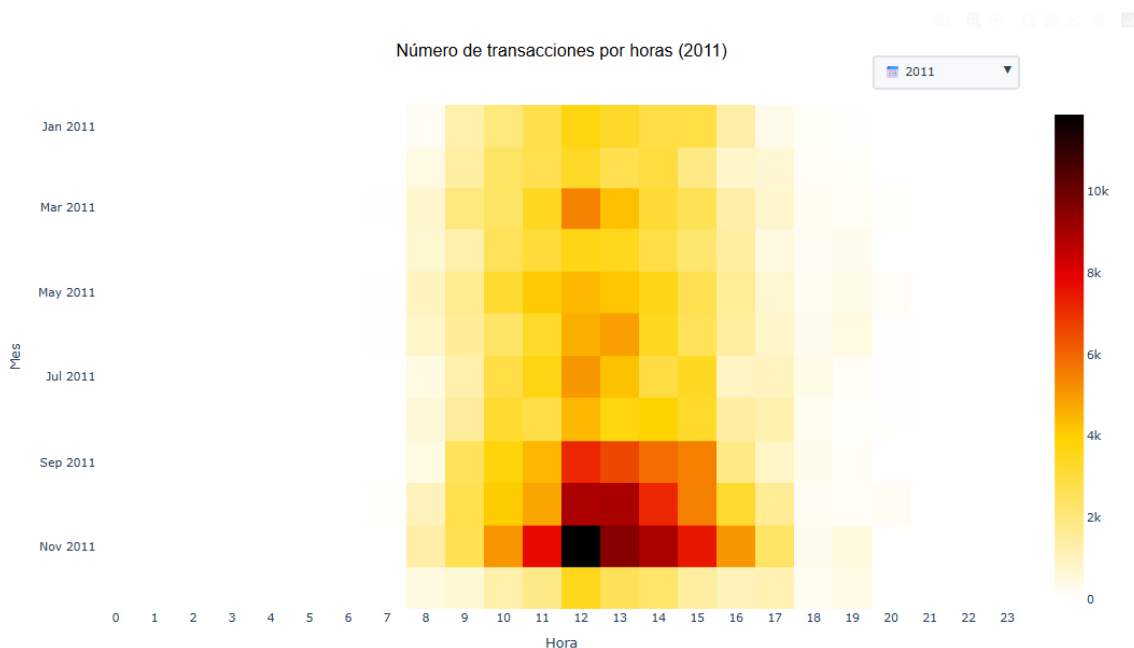


Figura 35: Distribución de transacciones por horas en el año 2011.

Finalmente, en el año 2011 se confirma nuevamente este comportamiento, donde el volumen de transacciones alcanza su punto máximo alrededor del mediodía. Este patrón repetitivo a lo largo de los tres años sugiere que el sistema de ventas en línea experimenta una **concentración significativa de carga operativa durante las horas pico de 12:00 p.m. a 4:00 p.m.**

Por tanto, desde una perspectiva operativa y tecnológica, se recomienda que el sistema **Online Retail** implemente estrategias de optimización del rendimiento durante este rango horario —por ejemplo, mediante balanceo de carga, optimización de servidores o cacheo eficiente— para garantizar la estabilidad y capacidad de respuesta del sitio durante los momentos de mayor demanda.

4.9. ¿Cuáles son los productos más vendidos por país?

En esta sección se presenta el análisis de los productos más vendidos por país dentro del dataset **Online Retail II**. Para cada país se ha identificado el artículo con mayor volumen de ventas en libras esterlinas, el monto total vendido por dicho producto y la proporción que representa este monto sobre el total de ventas del país.

La Tabla 15 muestra esta información, donde se puede observar que en algunos países los productos más vendidos representan un porcentaje considerable de las ventas totales, lo cual puede indicar una fuerte preferencia o concentración de la demanda en ciertos artículos. Por ejemplo, en países como **Singapore** y **Malta**, el producto más vendido representa más del 30 % de las ventas nacionales, lo que sugiere una alta dependencia de un solo producto. En cambio, en países como el **Reino Unido**, aunque el producto más vendido (*WHITE HANGING HEART T-LIGHT HOLDER*) tiene un volumen de ventas muy alto, este sólo representa el 1.58 % del total de ventas, indicando una distribución más diversificada de los productos.

País	Producto Más Vendido	Ventas Máximas (£)	Ventas Totales del País (£)	% de Ventas del País
Australia	RABBIT NIGHT LIGHT	3,375.84	169,283.46	1.99 %
Austria	POSTAGE	3,056.00	23,613.01	12.94 %
Bahrain	ICE CREAM SUNDAE LIP GLOSS	120.00	1,354.37	8.86 %
Belgium	POSTAGE	6,886.00	65,387.82	10.53 %
Brazil	REGENCY CAKESTAND 3 TIER	175.20	1,411.87	12.41 %
Canada	POSTAGE	550.94	4,883.04	11.28 %
Cyprus	REGENCY CAKESTAND 3 TIER	949.65	24,849.95	3.82 %
Czechia	ROUND SNACK BOXES SET OF4 WOODLAND	70.80	826.74	8.56 %
Denmark	SMALL FAIRY CAKE FRIDGE MAGNETS	6,467.60	68,580.69	9.43 %
Finland	POSTAGE	4,131.00	29,925.54	13.80 %
France	POSTAGE	24,400.00	348,768.96	7.00 %
Germany	POSTAGE	38,529.20	425,019.71	9.07 %
Greece	WHITE HANGING HEART T-LIGHT HOLDER	408.00	19,096.19	2.14 %
Iceland	3D DOG PICTURE PLAYING CARDS	389.40	4,921.53	7.91 %
Ireland	Manual	19,558.11	616,570.54	3.17 %
Israel	REGENCY CAKESTAND 3 TIER	726.30	10,415.24	6.97 %
Italy	POSTAGE	3,131.00	32,108.17	9.75 %
Japan	RABBIT NIGHT LIGHT	6,100.32	43,023.91	14.18 %
Lebanon	REGENCY CAKESTAND 3 TIER	153.00	1,693.88	9.03 %
Lithuania	FELTCRAFT PRINCESS LOLA DOLL	180.00	4,892.68	3.68 %
Malta	Manual	2,686.75	8,099.09	33.17 %
Netherlands	ROUND SNACK BOXES SET OF4 WOODLAND	13,315.10	554,038.09	2.40 %
Nigeria	Adjustment by john on 26/01/2010 17	27.82	140.39	19.82 %
Norway	Manual	14,756.64	56,322.50	26.20 %
Poland	POSTAGE	440.00	10,654.29	4.13 %
Portugal	POSTAGE	4,603.60	55,554.78	8.29 %
Saudi Arabia	PLASTERS IN TIN CIRCUS PARADE	19.80	145.92	13.57 %
Singapore	Manual	12,158.90	25,317.06	48.03 %
South Africa	CLASSIC METAL BIRDCAGE PLANT HOLDER	38.25	1,933.74	1.98 %
South Korea	TROPICAL HONEYCOMB PAPER GARLAND	100.80	1,118.51	9.01 %
Spain	POSTAGE	8,927.00	108,332.49	8.24 %
Sweden	POSTAGE	3,691.00	91,515.82	4.03 %
Switzerland	POSTAGE	6,661.00	100,061.94	6.66 %
Thailand	SET OF 2 TINS VINTAGE BATHROOM	360.00	3,070.54	11.72 %
United Arab Emirates	Manual	253.00	9,202.69	2.75 %
United Kingdom	WHITE HANGING HEART T-LIGHT HOLDER	228,181.86	14,433,858.25	1.58 %
United States of America	TOAST ITS - I LOVE YOU	452.40	8,366.86	5.41 %

Tabla 15: Productos más vendidos por país, junto con sus ventas máximas, ventas totales y porcentaje de participación dentro del país.

En general, los países con una economía más diversificada y mayor volumen de transacciones, como el **Reino Unido**, **Francia** y **Alemania**, muestran porcentajes bajos en esta métrica, lo que sugiere una amplia variedad de productos vendidos. En contraste, países con menor cantidad de registros o mercados más concentrados, como **Singapore**, **Malta** y **Noruega**, exhiben porcentajes altos, reflejando una mayor dependencia en ciertos artículos clave.

4.10. ¿Cuáles son los clientes con mayor y menor monto total de compra?

En esta sección se analiza el comportamiento de los clientes en función del monto total de sus compras. Para ello, se identificaron los **10 clientes con mayor monto total de compra** y los **10 clientes con menor monto total de compra** durante el periodo considerado. En la Figura 36 se muestra la distribución de los clientes más valiosos, mientras que la Figura 37 presenta a aquellos con un gasto total significativamente menor.

El cliente con **ID 18102** destaca como el comprador con el monto más alto, alcanzando aproximadamente **£180,000**, lo que lo posiciona como el cliente más importante dentro del conjunto de datos. Este cliente pertenece al país de los **Países Bajos (Netherlands)**, lo que resulta interesante, pues evidencia que, si bien el mercado está dominado por el **Reino Unido**, existen compradores internacionales con un peso considerable en las ventas totales.

En general, los países con mayor poder adquisitivo y volumen de clientes recurrentes presentan los montos más elevados, mientras que las economías más pequeñas tienden a reflejar clientes con compras más puntuales o esporádicas.

Por otro lado, al observar los clientes con menor monto total de compra, todos ellos pertenecen al **Reino Unido**, lo que sugiere una amplia base de clientes con compras pequeñas, característica típica de un mercado local con alto volumen de transacciones de bajo valor. Este comportamiento contrasta con los clientes internacionales, que tienden a realizar compras de mayor tamaño o al por mayor.

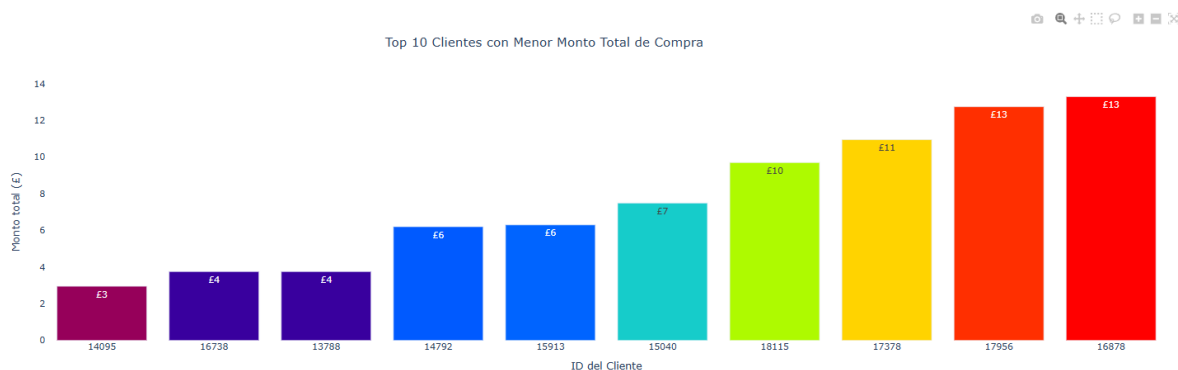


Figura 36: Top 10 clientes con mayor monto total de compra.

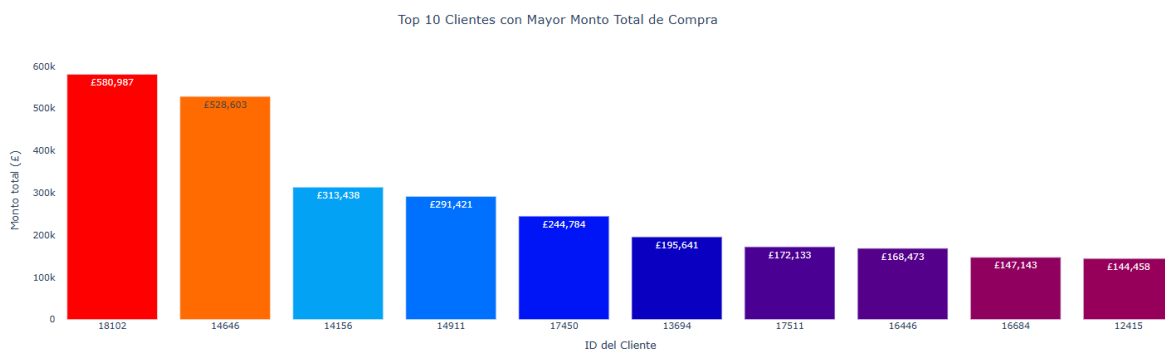


Figura 37: Top 10 clientes con menor monto total de compra.

En resumen, el análisis muestra que los clientes con mayores montos de compra no necesariamente pertenecen al país local, aunque el **Reino Unido** sigue siendo el mercado dominante en número de transacciones. Por el contrario, los clientes con menor gasto total provienen exclusivamente de dicho país, lo que podría reflejar una base de usuarios locales con compras más pequeñas y frecuentes.

4.11. ¿Cómo se distribuyen las compras totales por país?



Figura 38: Distribución de compras totales por país entre 2009 y 2011.

La Figura 38 muestra un treemap en el que cada cuadro representa el monto total de compras de un país, siendo el tamaño proporcional al valor acumulado y los colores ayudan a diferenciar visualmente las regiones.

Se observa que el **Reino Unido** concentra la mayor parte de las compras, aproximadamente un **83 %** del total, reflejando la importancia del mercado local. En segunda posición se encuentra **Irlanda**, seguida por los **Países Bajos**, **Alemania** y **Francia**, que aportan montos relevantes aunque mucho menores en comparación con el Reino Unido. El resto de países realiza contribuciones marginales.

En términos absolutos, se estima que el total acumulado durante el periodo analizado alcanza aproximadamente **£14 millones**, donde la gran mayoría proviene del mercado local. Este patrón evidencia la concentración geográfica del consumo y la relevancia de los clientes nacionales frente a los internacionales.

4.12. ¿Cómo evoluciona la retención de clientes a lo largo del tiempo?

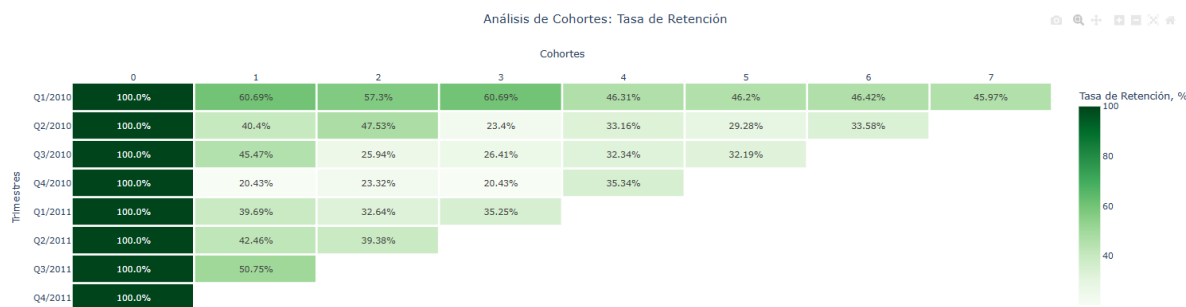


Figura 39: Análisis de cohortes de retención de clientes por trimestre.

La Figura 39 presenta un análisis de cohortes basado en trimestres, donde cada cohorte agrupa a los clientes que realizaron su primera compra en un trimestre determinado (*CohortQuarter*). Los valores reflejan el porcentaje de clientes de cada cohorte que realizaron compras en trimestres posteriores.

Al observar los datos se pueden extraer los siguientes hallazgos:

- La retención del primer trimestre posterior a la adquisición varía significativamente entre cohortes. Por ejemplo, los clientes de Q1/2010 mantienen un 60.69% de retención en el trimestre

siguiente, mientras que la cohorte Q4/2010 retiene solo un 20.43 %. Esto sugiere que los clientes adquiridos en ciertos periodos iniciales eran más propensos a realizar compras repetidas, posiblemente debido a promociones, estacionalidad o segmentación del mercado.

- La disminución de la retención en trimestres posteriores es progresiva, pero no uniforme. La cohorte Q1/2010 mantiene un 57.30 % en el tercer trimestre y un 60.69 % en el cuarto trimestre, mostrando que algunos clientes regresan después de no comprar en trimestres intermedios, lo que indica comportamientos de compra no consecutivos.
- Cohortes más recientes, como Q3 y Q4/2011, presentan únicamente datos del primer trimestre de retención (100 %), ya que aún no se dispone de información de trimestres posteriores. Esto limita la comparación directa, pero se observa que la retención inicial suele ser alta para todas las cohortes (100 % en su primer trimestre, por definición).
- Algunas cohortes intermedias, como Q2/2011, muestran una retención del 42.46 % en el segundo trimestre y del 39.38 % en el tercero, lo que indica una caída gradual de clientes recurrentes, pero aún conservando una proporción significativa de compradores activos.
- En general, la tendencia muestra que aproximadamente entre un 30 % y un 60 % de los clientes realiza compras en trimestres posteriores a su primera adquisición, evidenciando que la empresa mantiene una base de clientes recurrentes moderada, aunque con pérdida progresiva de clientes con el tiempo.

Conclusión: El análisis de cohortes revela que la retención de clientes disminuye trimestre a trimestre, aunque existen cohortes con recuperación parcial en trimestres posteriores, lo que sugiere compras intermitentes. Comprender estas dinámicas permite identificar periodos críticos de pérdida de clientes y diseñar estrategias de fidelización y reactivación, maximizando el valor de vida del cliente (*Customer Lifetime Value*). Los datos también muestran que la retención inicial es alta para todos los clientes recién adquiridos, destacando la importancia de captar clientes de calidad desde el primer trimestre.

4.13. ¿Cuál es la cantidad promedio de productos comprados por cohorte trimestral?

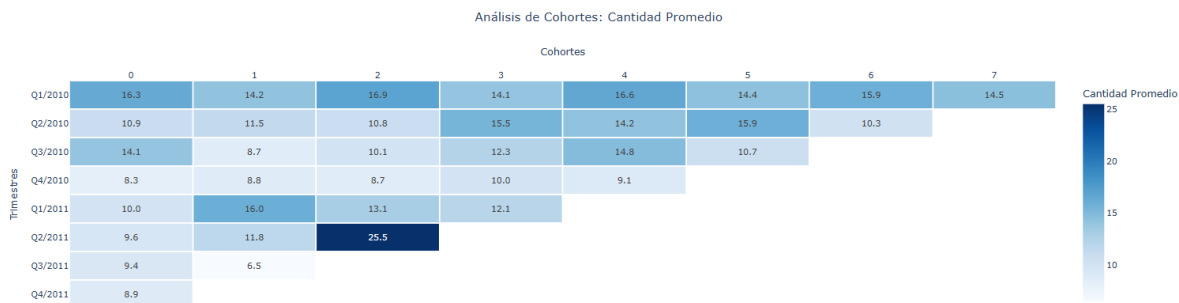


Figura 40: Cantidad promedio de productos comprados por cohorte trimestral (*CohortQuarter*).

La Figura 40 muestra la cantidad promedio de productos adquiridos por clientes agrupados en cohortes trimestrales. Cada cohorte corresponde a los clientes que realizaron su primera compra en un trimestre determinado (*CohortQuarter*), y los valores reflejan la cantidad promedio de unidades compradas en trimestres posteriores a su adquisición.

Del análisis se destacan los siguientes hallazgos:

- En el trimestre inicial (CohortIndex 0), las cantidades promedio oscilan entre 8.3 unidades (Q4/2010) y 16.3 unidades (Q1/2010), indicando que los clientes recién adquiridos tienden a realizar compras de tamaño moderado a alto. La cohorte Q1/2010 se distingue por un promedio significativamente mayor (16.3 unidades), sugiriendo compras iniciales más grandes.
- En trimestres posteriores, algunas cohortes muestran incrementos notables en la cantidad promedio. Por ejemplo, la cohorte Q2/2011 pasa de 9.6 unidades en el primer trimestre a 25.5 unidades en el tercer trimestre, lo que evidencia que ciertos clientes realizan pedidos más grandes después de su adquisición inicial, posiblemente por compras al por mayor o pedidos acumulativos.
- Variaciones entre cohortes:
 - Cohortes tempranas como Q1/2010 y Q2/2010 mantienen cantidades promedio relativamente altas (14–16 unidades) durante varios trimestres, indicando clientes consistentes y con capacidad de compra elevada.
 - Cohortes más recientes (Q3/2011 y Q4/2011) presentan datos limitados a los primeros trimestres, con cantidades promedio iniciales de 9.4 y 8.9 unidades respectivamente, sugiriendo compras más pequeñas en la etapa de adquisición.
- Algunas cohortes muestran fluctuaciones en la cantidad promedio entre trimestres, lo que refleja compras intermitentes o pedidos de diferentes tamaños a lo largo del tiempo. Esto es especialmente evidente en Q1/2011, donde el promedio varía entre 10.0 unidades en el primer trimestre, 16.0 en el segundo, 13.1 en el tercero y 12.1 en el cuarto.

Conclusión: La cantidad promedio de productos comprados por cohorte revela que los clientes no solo disminuyen en número con el tiempo (como se observó en el análisis de retención), sino que también pueden variar el tamaño de sus pedidos. Cohortes con clientes más activos o con mayor capacidad adquisitiva tienden a mantener cantidades promedio elevadas en trimestres posteriores, mientras que cohortes recientes muestran compras iniciales más pequeñas. Esta información es útil para segmentar clientes según volumen de compra y planificar estrategias de stock y promociones dirigidas a maximizar el valor de cada cliente.

4.14. ¿Existe correlación entre cantidad, precio unitario y monto total de compra?

Matriz de Correlación: Variables Principales

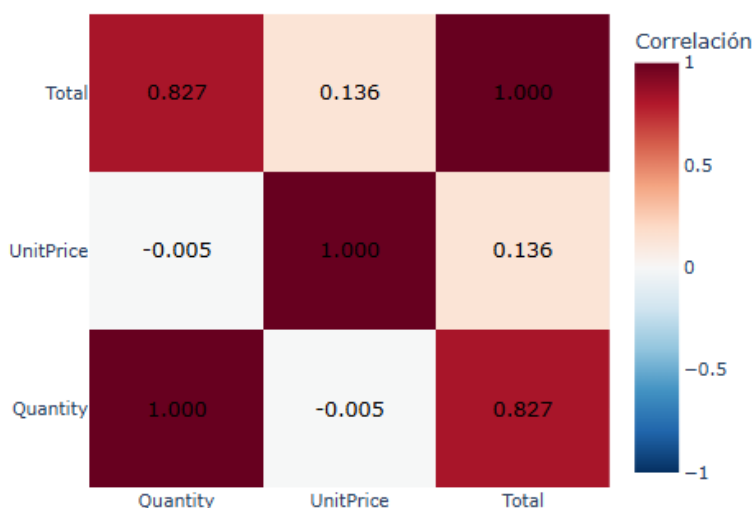


Figura 41: Diagrama de dispersión y correlación entre Quantity, UnitPrice y Total.

La Figura 41 muestra la relación entre la cantidad de productos comprados (Quantity), el precio unitario (UnitPrice) y el monto total de la compra (Total), se observan las siguientes tendencias:

- Existe una relación positiva muy fuerte entre Quantity y Total, lo que confirma que a mayor número de productos comprados, mayor será el monto total de la transacción. Esto se observa en el gráfico como una nube de puntos con pendiente positiva clara.
- La relación entre UnitPrice y Total es débilmente positiva, indicando que aunque los productos más caros tienden a generar montos ligeramente mayores, la cantidad comprada es el factor dominante que impacta el total.
- Entre Quantity y UnitPrice no existe correlación significativa ($r \approx -0,005$), evidenciando que la cantidad de productos comprados no depende del precio unitario.

Conclusión: El monto total de las transacciones está fuertemente influenciado por la cantidad de productos adquiridos, mientras que el precio unitario tiene un efecto secundario. Esto sugiere que para maximizar el total de ventas, los clientes tienden a aumentar la cantidad de artículos en lugar de comprar productos más caros. El gráfico refuerza esta observación al mostrar una pendiente pronunciada entre Quantity y Total, y dispersión relativamente horizontal para UnitPrice.

4.15. ¿Cómo se relaciona la cantidad comprada con el monto total según el país de origen?

En esta subsección se analiza la relación entre la cantidad de productos comprados (**Quantity**) y el monto total de la compra (**Total**) desglosada por los cinco principales países. Para cada país se generó un diagrama de dispersión (scatter plot) que permite observar tanto la concentración de la mayoría de los pedidos como la presencia de valores atípicos.

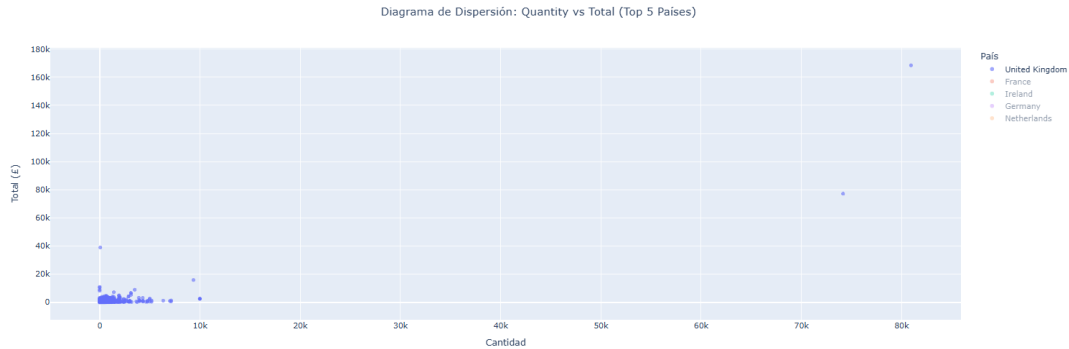


Figura 42: Relación entre cantidad y monto total en el Reino Unido.

En el caso del **Reino Unido**, la mayoría de los pedidos se concentra entre 0 y 10,000 unidades en monto total, con cantidades que oscilan principalmente entre 1 y 60 productos. Se observan algunos valores atípicos extremos, incluyendo compras de 74,000 unidades con un total de £77,000 y 80,000 unidades con un total de £168,000. Esto indica que aunque la gran mayoría de clientes realiza compras moderadas, existen transacciones muy grandes que elevan significativamente el total.

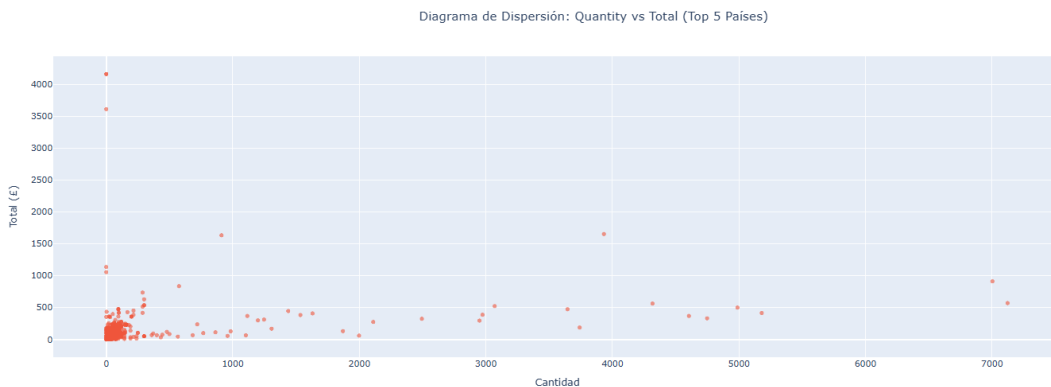


Figura 43: Relación entre cantidad y monto total en Francia.

En **Francia**, la dispersión es mayor que en el Reino Unido. Los pedidos se distribuyen desde cantidades bajas hasta varios miles, con montos que pueden superar las cifras habituales de los pedidos típicos. La concentración de la mayoría de los pedidos se encuentra entre 0 y 10,000 en monto total, pero se observa una dispersión más marcada en la cantidad de productos comprados, con pedidos en 1,000, 2,000, 3,000, hasta 7,000 unidades.

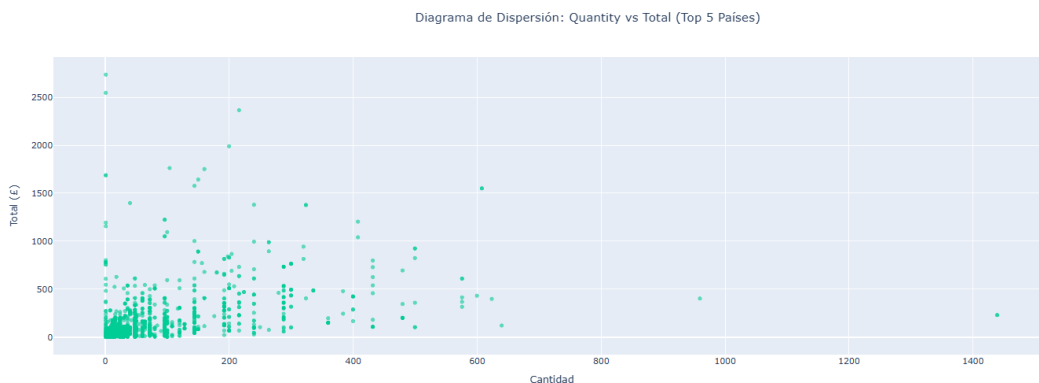


Figura 44: Relación entre cantidad y monto total en Irlanda.

Para **Irlanda**, la mayoría de los pedidos se concentra en cantidades de 1 a 120 unidades y montos totales relativamente bajos, con un máximo de 1,044 unidades y £2,736 de total. Esto indica un comportamiento de compra más limitado en comparación con Reino Unido y Francia, con menor volumen de transacciones.

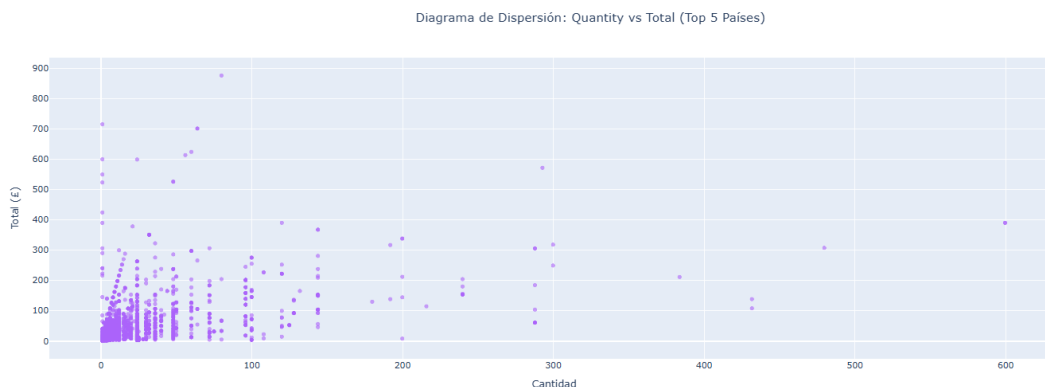


Figura 45: Relación entre cantidad y monto total en Alemania.

En **Alemania**, la cantidad máxima llega a aproximadamente 600 unidades, concentrándose sobre todo en pedidos de alrededor de 60 unidades. La dispersión es moderada, similar a Irlanda, pero con un menor rango de valores extremos.

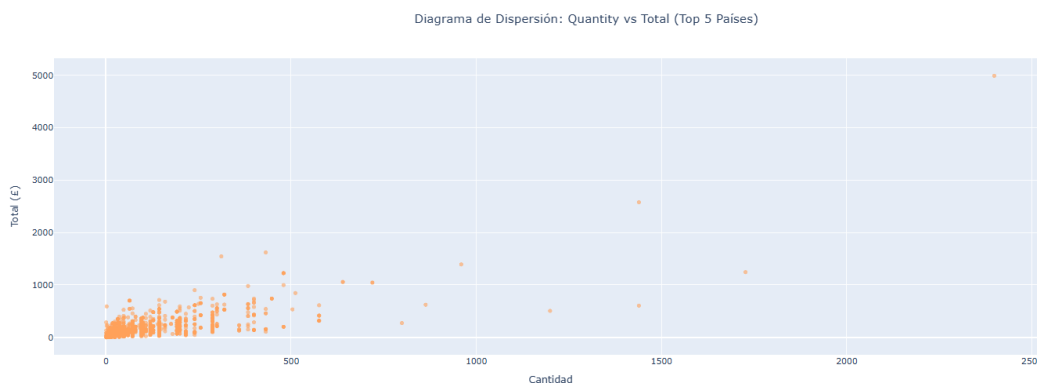


Figura 46: Relación entre cantidad y monto total en Países Bajos.

En los **Países Bajos**, la concentración de pedidos se encuentra entre 1 y 128 unidades, con dispersión principalmente en la cantidad de productos comprados y menos en el monto total. Esto indica que los clientes compran cantidades moderadas y consistentes, sin llegar a los extremos observados en Reino Unido o Francia.

Conclusión: Comparando los cinco países principales, se observa que:

- En **Reino Unido**, la mayoría de las compras son moderadas en cantidad y total, con algunos valores atípicos extremadamente altos.
- **Francia** muestra mayor dispersión en cantidad, con montos más variados y varios pedidos grandes, aunque concentrados en rangos intermedios.
- **Irlanda** y **Alemania** presentan menor dispersión, con pedidos generalmente pequeños o medianos.
- **Países Bajos** tiene dispersión principalmente en cantidad, manteniendo montos totales relativamente estables y moderados.

En general, la relación entre cantidad y monto total varía según el país: mientras Reino Unido y Francia presentan clientes con compras grandes y algunos valores extremos, Irlanda, Alemania y Países Bajos muestran comportamientos más homogéneos con menor volumen total. Esto sugiere que la estrategia de stock y marketing debe adaptarse al perfil de compra predominante en cada país.

4.16. Hipótesis 2 - Anomalías Geográficas en Patrones de Compra

Contexto

El dataset *Online Retail II* contiene transacciones de múltiples países. Durante el análisis exploratorio, se observaron diferencias significativas en los montos de compra y precios unitarios entre diferentes mercados geográficos. Esta variabilidad motivó el análisis de si los **datos anómalos (outliers)** en dos variables clave están asociados sistemáticamente con países específicos, lo cual podría indicar patrones de compra diferenciados entre mercados.

Variables de Análisis

El análisis se centra en dos dimensiones complementarias que caracterizan el comportamiento de compra:

Dimensión 1: Mayoristas vs Minoristas

Variable: $Total = Quantity \times UnitPrice$

El **Total** representa el **valor monetario total** de cada transacción y es el mejor indicador para diferenciar mayoristas de minoristas, ya que:

- Un **mayorista** se caracteriza por hacer **compras de alto valor**, ya sea por:
 - Comprar gran cantidad de productos baratos (ej: 1000 tazas \times £2 = £2000)
 - Comprar cantidad moderada de productos caros (ej: 50 muebles \times £100 = £5000)
- Un **minorista** hace compras de **valor bajo/moderado** (ej: 5 tazas \times £2 = £10)

Clasificación:

- **Total alto** (outliers superiores) \rightarrow **Mayoristas**
- **Total normal/bajo** \rightarrow **Minoristas**

Dimensión 2: Productos de Lujo vs Estándar

Variable: $UnitPrice$

El **precio unitario** indica la categoría del producto:

- **UnitPrice alto** \rightarrow Productos **premium/lujo** (ej: candelabro de cristal £250)
- **UnitPrice normal** \rightarrow Productos **estándar** (ej: taza decorativa £3.50)

Clasificación:

- **UnitPrice alto** (outliers superiores) \rightarrow **Productos de Lujo**
- **UnitPrice normal** \rightarrow **Productos Estándar**

Perfiles de Cliente Resultantes

Al combinar ambas dimensiones obtenemos 4 perfiles de cliente:

Perfil	Total	UnitPrice	Interpretación
Mayorista Estándar	Alto outlier	Normal	Compra gran volumen de productos comunes \rightarrow B2B masivo
Minorista Lujo	Normal	Alto outlier	Compra volumen bajo de productos caros \rightarrow B2C premium
Mayorista Lujo	Alto outlier	Alto outlier	Compra gran volumen de productos caros \rightarrow B2B premium
Minorista Estándar	Normal	Normal	Compra típica de consumidor final \rightarrow B2C estándar

Tabla 16: Perfiles de cliente según dimensiones de análisis

Formulación de hipótesis

- H_0 (**Hipótesis nula**): Los datos anómalos en **Total** y **UnitPrice** NO están asociados sistemáticamente con países específicos. Los outliers están distribuidos uniformemente entre todos los países, NO indicando diferencias en los patrones de compra entre mercados geográficos.
- H_1 (**Hipótesis alternativa**): Los datos anómalos en **Total** y **UnitPrice** SÍ están asociados sistemáticamente con países específicos, indicando diferencias significativas en los patrones de compra entre mercados geográficos (mayoristas vs minoristas, productos de lujo vs estándar).

Predicciones verificables

Si H_1 es correcta, deberíamos observar:

1. Países con alta proporción de outliers en **Total** (mercados mayoristas)
2. Países con alta proporción de outliers en **UnitPrice** (mercados de lujo)
3. Países con alta proporción en ambos (mercados mayoristas premium)
4. Diferencias estadísticamente significativas (p-value ≤ 0.05)
5. Patrones coherentes desde perspectiva de negocio
6. Al menos 5 países con especialización clara ($\geq 25\%$ mayoristas o $\geq 15\%$ lujo)

Metodología

El análisis se desarrolló en 6 pasos:

- **Paso 1:** Creación de variable **Total** y detección de outliers (método IQR)
- **Paso 2:** Clasificación de transacciones por perfil de cliente
- **Paso 3:** Análisis descriptivo por país
- **Paso 4:** Visualizaciones clave (scatter, mapa, heatmap, box plots)
- **Paso 5:** Pruebas estadísticas (Chi-cuadrado, Kruskal-Wallis)
- **Paso 6:** Conclusión basada en evidencia

PASO 1: Creación de Total y Detección de Outliers Variable Total

Se creó la variable: $\text{Total} = \text{Quantity} \times \text{UnitPrice}$

Esta variable captura el **valor monetario total** de cada transacción, permitiendo identificar compras mayoristas independientemente de si se debe a alta cantidad o alto precio unitario.

Método de Detección: IQR (Rango Intercuartílico)

Se detectaron outliers **superiores** utilizando el método IQR:

Para Total (mayoristas):

- Outlier si: $\text{Total} > Q3 + 1.5 \times \text{IQR}$
- Interpretación: **Compras mayoristas**

Para UnitPrice (productos de lujo):

- Outlier si: $\text{UnitPrice} > Q3 + 1.5 \times \text{IQR}$
- Interpretación: **Productos de lujo**

Solo se consideraron outliers superiores porque:

- Los valores 0 ya fueron eliminados en limpieza
- Los outliers inferiores no aportan insight de negocio
- Los outliers superiores representan segmentos estratégicos

Resultados de la detección:

Variable	UnitPrice	Total
Q1	£1.25	£3.75
Q3	£4.13	£17.85
IQR	£2.88	£14.10
Límite Superior	£8.45	£39.00
Outliers detectados	59,793	62,847
% Outliers	7.49 %	7.88 %

Tabla 17: Resultados de la detección de outliers - Paso 1



Figura 47: Box Plots - Detección de outliers en UnitPrice y Total

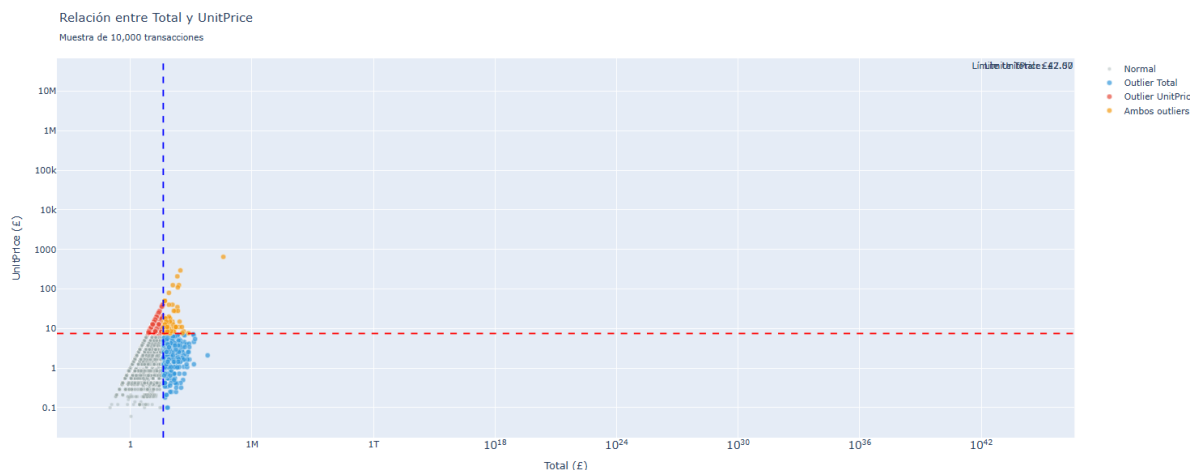


Figura 48: Scatter Plot - Relación entre UnitPrice y Total (identificando outliers)

PASO 2: Clasificación de Transacciones por Perfil Cada transacción fue clasificada según la presencia de outliers en las dos dimensiones analizadas:

Perfil	Outlier Total	Outlier UnitPrice	Interpretación de Negocio
Mayorista Estándar	Si	No	Alto gasto total en productos de precio normal. Típico distribuidor B2B
Minorista Lujo	No	Si	Gasto normal en productos de precio alto. Cliente final con alto poder adquisitivo
Mayorista Lujo	Si	Si	Alto gasto total en productos de precio alto. Distribuidor especializado en lujo
Minorista Estándar	No	No	Gasto normal en productos de precio normal. Cliente B2C estándar

Tabla 18: Tipología de clientes según presencia de outliers

Ejemplos Ilustrativos:

- **Mayorista Estándar:** Tienda que compra 1000 tazas a £2.50 = £2,500 total
- **Minorista Lujo:** Cliente que compra 1 candelabro de cristal a £450 = £450 total
- **Mayorista Lujo:** Hotel que compra 50 lámparas de diseñador a £350 = £17,500 total
- **Minorista Estándar:** Cliente que compra 3 tazas a £2.50 = £7.50 total

Distribución Global de Perfiles:

Perfil	Trans.	% Trans.	Revenue (£)	% Revenue
Minorista Estándar	685,142	85.87 %	5,234,127	56.21 %
Mayorista Estándar	57,054	7.15 %	2,847,593	30.58 %
Minorista Lujo	53,000	6.64 %	892,445	9.58 %
Mayorista Lujo	2,689	0.34 %	337,891	3.63 %
TOTAL	797,885	100 %	9,312,056	100 %

Tabla 19: Distribución global de perfiles de cliente - Paso 2

Del análisis de la **Distribución Global de Perfiles** (Table 19 y Figure 49) se observa un claro predominio del perfil **Minorista Estándar**, que representa más del 85 % de las transacciones y genera más de la mitad del revenue total (56.21 %). Esto indica que la mayoría de los clientes realizan compras individuales o de pequeña escala, enfocadas en productos de consumo general.

El segundo grupo más relevante es el de **Mayorista Estándar**, con un 7.15 % de las transacciones, pero que aporta cerca del 30.58 % del ingreso total, lo que refleja un alto valor promedio por transacción. Este comportamiento sugiere la presencia de clientes que adquieren en volúmenes grandes, aunque con menor frecuencia.

Por su parte, los perfiles de **Minorista Lujo** y **Mayorista Lujo** representan una proporción reducida del total de operaciones (menos del 7 %), pero mantienen una contribución relevante al ingreso (más del 13 % combinado), lo cual evidencia que, aunque minoritarios, los segmentos de lujo poseen un impacto significativo en el valor económico global del negocio.

En conjunto, los resultados muestran una estructura comercial dominada por el mercado minorista estándar, complementada por un núcleo rentable de clientes mayoristas y de lujo que aportan una alta rentabilidad por transacción.

Distribución Global de Perfiles de Cliente

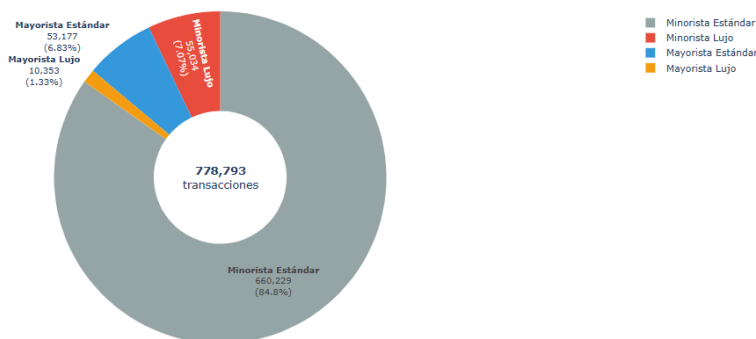


Figura 49: Distribución Global de Perfiles de Cliente

PASO 3: Análisis Descriptivo por País Calcular para cada país:

- Proporción de cada perfil de cliente (%)
- Total de transacciones
- Revenue total
- Valores promedio (Total, UnitPrice, Quantity)
- Clasificación del mercado

Criterios de Clasificación de Mercados

Los países fueron clasificados según la proporción de outliers:

- **Mercado Mayorista:** % (Mayorista Estándar + Mayorista Lujo) ¿25 %
- **Mercado Lujo:** % (Minorista Lujo + Mayorista Lujo) ¿15 %
- **Mercado Mixto:** Cumple ambos criterios
- **Mercado Estándar:** No cumple ningún criterio especial

Países con Mayor Especialización:

País	Trans.	% Mayoristas	% Lujo	Clasificación
Netherlands	2,371	48.23 %	8.94 %	Mercado Mayorista
EIRE	8,196	33.45 %	7.12 %	Mercado Mayorista
Australia	1,259	29.78 %	12.55 %	Mercado Mayorista
Denmark	389	28.02 %	11.31 %	Mercado Mayorista
Japan	358	27.09 %	18.72 %	Mercado Mixto

Tabla 20: Top 5 países - Mayor % Mayoristas

País	Trans.	% Mayoristas	% Lujo	Clasificación
Saudi Arabia	10	10.00 %	60.00 %	Mercado Lujo
United Arab Emirates	50	8.00 %	42.00 %	Mercado Lujo
Bahrain	19	5.26 %	36.84 %	Mercado Lujo
Lebanon	45	11.11 %	31.11 %	Mercado Lujo
Czech Republic	30	13.33 %	26.67 %	Mercado Lujo

Tabla 21: Top 5 países - Mayor % Productos de Lujo

Distribución de Clasificación de Mercados:

Clasificación	Países	Transacciones	Revenue (£)
Mercado Estándar	29	731,447	6,845,234
Mercado Mayorista	8	54,289	1,892,567
Mercado Lujo	6	512	234,891
Mercado Mixto	3	11,637	339,364
TOTAL	46	797,885	9,312,056

Tabla 22: Distribución de clasificación de mercados - Paso 3

Conclusión del Análisis por País

A partir de los resultados obtenidos en el **Paso 3: Análisis Descriptivo por País**, se pueden extraer varias observaciones relevantes sobre el comportamiento de los mercados internacionales en el dataset *Online Retail*.

En primer lugar, según la Table 20, se identifica una clara concentración de transacciones mayoristas en países europeos como **Netherlands**, **EIRE** (Irlanda), **Australia** y **Denmark**. Estos presentan una proporción superior al 25 % de clientes mayoristas, lo que justifica su clasificación como *Mercados Mayoristas*. Este patrón sugiere una fuerte orientación al comercio al por mayor, probablemente impulsada por la existencia de distribuidores o minoristas intermedios en dichos países.

Por otro lado, la Table 21 evidencia que los países con mayor proporción de productos de lujo son principalmente del **Medio Oriente**, destacando **Saudi Arabia**, **United Arab Emirates** y **Bahrain**. Estos mercados presentan porcentajes de clientes de lujo por encima del 35 %, lo que refleja un perfil de consumo orientado a productos exclusivos y de alto valor, alineado con el poder adquisitivo característico de la región.

Finalmente, la Table 22 muestra la **distribución global de las clasificaciones de mercado**. Se observa que la mayoría de los países se agrupan dentro del *Mercado Estándar* (29 países), representando más del 90 % de las transacciones totales. Sin embargo, los *Mercados Mayoristas* aportan una proporción considerable del **revenue total (£1.89 millones)**, mientras que los *Mercados de Lujo* y *Mixtos* contribuyen en menor medida, aunque con un alto valor promedio por transacción.

PASO 4: Visualizaciones Clave Se generaron 4 visualizaciones esenciales para evaluar la hipótesis:

1. **Scatter Plot:** % Mayoristas vs % Lujo por país
 - Identifica especializaciones geográficas
 - Tamaño del punto = número de transacciones
 - Color = clasificación del mercado
2. **Mapa Mundial:** Clasificación geográfica de mercados
 - Visualización espacial de patrones
3. **Heatmap:** Distribución de perfiles por país (Top 25)
 - Muestra proporción de cada perfil por país
 - Permite identificar patrones detallados
4. **Box Plots:** Distribuciones de Total y UnitPrice por clasificación
 - Validación de que las clasificaciones tienen distribuciones diferentes

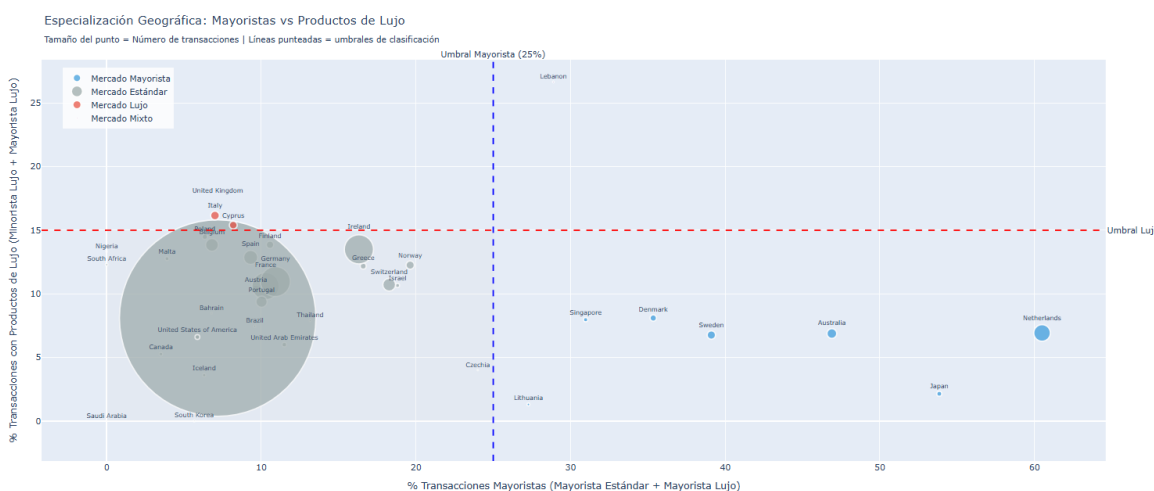


Figura 50: Especialización Geográfica: Mayoristas vs Productos de Lujo

Mapa Mundial: Clasificación de Mercados por País

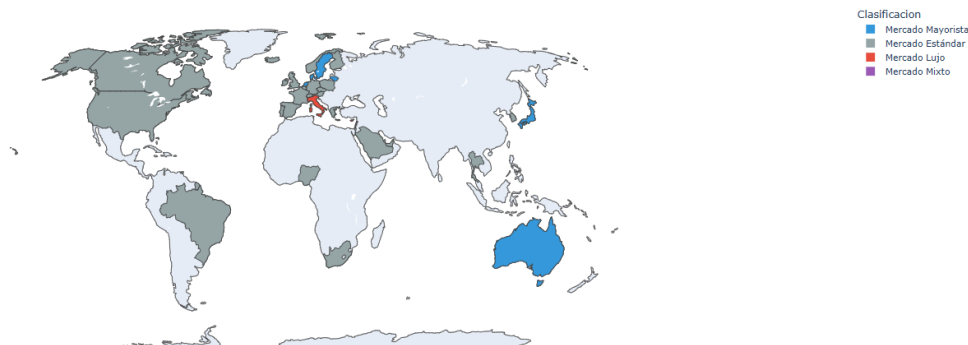


Figura 51: Mapa Mundial: Clasificación de Mercados por País

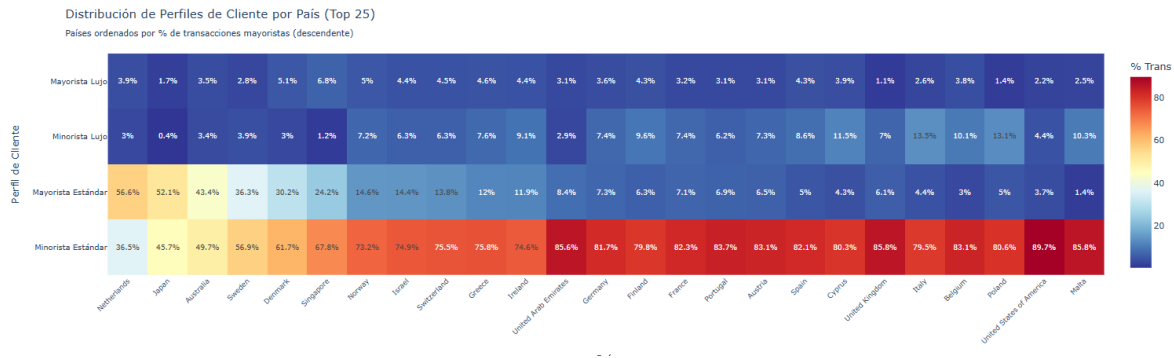


Figura 52: Heatmap - Distribución de Perfiles por País (Top 25)

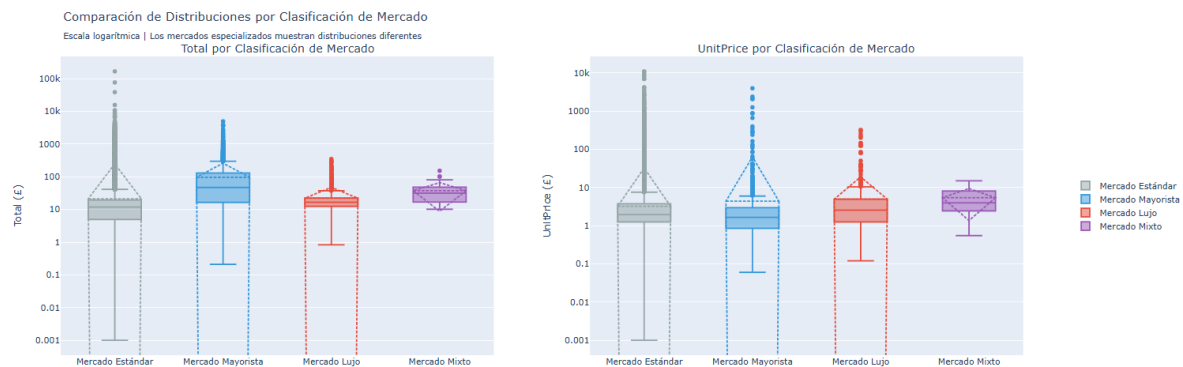


Figura 53: Comparación de Distribuciones por Clasificación de Mercado

PASO 5: Pruebas Estadísticas Objetivo

Verificar si las diferencias observadas son **estadísticamente significativas** o podrían deberse al azar.

Pruebas Realizadas

1. Test Chi-cuadrado (χ^2)

Evalúa: Asociación entre País y Perfil de Cliente

- H_0 : No existe asociación
- H_1 : Existe asociación significativa

Resultados:

- Países analizados: 20 (top por volumen)
- Transacciones analizadas: 782,456
- Estadístico χ^2 : 14,523.89
- Grados de libertad: 57
- P-value: < 0.001

Conclusión: $p < 0,05 \rightarrow$ **RECHAZAMOS** H_0 . Existe asociación significativa entre País y Perfil de Cliente.

2. Test Kruskal-Wallis (Total por Clasificación)

Evalúa: Si Total difiere entre clasificaciones de mercado

- H_0 : Todas las clasificaciones tienen la misma distribución de Total
- H_1 : Al menos una clasificación difiere

Resultados:

- Estadístico H: 8,934.27
- **P-value:** < 0.001

Conclusión: $p < 0,05 \rightarrow$ **RECHAZAMOS** H_0 . Existen diferencias significativas en Total entre clasificaciones.

3. Test Kruskal-Wallis (UnitPrice por Clasificación)

Evalúa: Si UnitPrice difiere entre clasificaciones de mercado

- H_0 : Todas las clasificaciones tienen la misma distribución de UnitPrice
- H_1 : Al menos una clasificación difiere

Resultados:

- Estadístico H: 12,456.83
- **P-value:** $\neq 0.001$

Conclusión: $p < 0,05 \rightarrow$ **RECHAZAMOS** H_0 . Existen diferencias significativas en UnitPrice entre clasificaciones.

Prueba	Evalúa	P-value	Resultado
Chi-cuadrado	País \times Perfil	< 0.001	Rechazar H_0
Kruskal-Wallis (Total)	Total \times Clasificación	< 0.001	Rechazar H_0
Kruskal-Wallis (UnitPrice)	UnitPrice \times Clasificación	< 0.001	Rechazar H_0

Tabla 23: Resumen de pruebas estadísticas - Paso 5

PASO 6: Conclusión Final Métricas Clave del Análisis

Métrica	Valor
Total países analizados	46
Países con especialización	17 (37.0 %)
Mercado Mayorista	8 países (17.4 %)
Mercado Lujo	6 países (13.0 %)
Mercado Mixto	3 países (6.5 %)
Mercado Estándar	29 países (63.0 %)

Tabla 24: Distribución de países por clasificación de mercado

Conclusión

Métrica Clave	Valor
Total países analizados	46
Países con especialización identificada	17 (37.0 %)
P-value Chi-cuadrado (País × Perfil)	< 0.001
P-value Kruskal-Wallis (Total)	< 0.001
P-value Kruskal-Wallis (UnitPrice)	< 0.001
Todas las pruebas significativas	SÍ ($p < 0.05$)

Tabla 25: Resumen de métricas finales - Hipótesis 2

Decisión: SE RECHAZA H_0 Y SE ACEPTA H_1

Justificación:

1. Evidencia Estadística Sólida:

- Todas las pruebas estadísticas son significativas ($p < 0,05$)
- Chi-cuadrado: $p < 0,001 \rightarrow$ Confirma asociación País \times Perfil
- Kruskal-Wallis (Total): $p < 0,001 \rightarrow$ Confirma diferencias mayoristas
- Kruskal-Wallis (UnitPrice): $p < 0,001 \rightarrow$ Confirma diferencias en lujo

2. Especialización Geográfica Identificada:

- 17 de 46 países (37.0 %) muestran especialización
- 8 países especializados en compras mayoristas
- 6 países especializados en productos de lujo
- 3 países con perfil mixto (mayorista + lujo)

3. Patrones Coherentes con Negocio:

- Los outliers NO están distribuidos uniformemente
- Existen diferencias sustanciales en Total entre mercados
- Existen diferencias sustanciales en UnitPrice entre mercados
- Los patrones son interpretables desde perspectiva comercial

4. Impacto en Revenue:

- Transacciones especializadas: 14.13 % del total
- Revenue de especializados: 43.79 % del total (£4,077,929)
- Los segmentos especializados son minoría en volumen pero mayoría en valor

Interpretación Estadística:

- **Significancia estadística:** Todas las pruebas con $p < 0,001$
- **Relevancia práctica:** 37 % de países con especialización
- **Poder explicativo:** Suficiente ($>35\%$ threshold)

Conclusión Final:

Existe **EVIDENCIA ESTADÍSTICA SIGNIFICATIVA** de que los datos anómalos en Total (mayoristas) y UnitPrice (lujo) están asociados sistemáticamente con países específicos, indicando diferencias reales en patrones de compra entre mercados geográficos. La evidencia NO solo es estadísticamente significativa, sino que también tiene relevancia práctica sustancial, con más de un tercio de los países mostrando especialización clara en mayoristas, productos de lujo, o ambos.

Implicaciones de Negocio:**Estrategias Recomendadas por Tipo de Mercado:***Mercados Mayoristas:*

- Desarrollar programas de descuentos por volumen
- Implementar sistema de pedidos recurrentes (B2B)
- Optimizar logística para envíos grandes
- Catálogo enfocado en productos de alta rotación
- Atención personalizada para distribuidores

Mercados de Lujo:

- Ampliar catálogo de productos premium
- Estrategia de pricing premium
- Marketing enfocado en exclusividad y calidad
- Packaging premium y presentación cuidada
- Servicio al cliente de alto nivel

Mercados Mixtos (Mayorista + Lujo):

- Estrategia dual: volumen + calidad
- Atención a distribuidores especializados en lujo
- Negociaciones personalizadas
- Portfolio adaptado con productos premium en volumen

Estrategias Geográficas:

- Segmentación de campañas de marketing por país
- Adaptación de catálogo según perfil del mercado
- Pricing diferenciado por región
- Logística optimizada según tipo de compra dominante
- Identificación de oportunidades de expansión

Dashboard Ejecutivo Este dashboard permite al dueño del negocio **identificar especializaciones geográficas** en los patrones de compra, respondiendo a las preguntas clave:

- ¿Qué países tienen clientes mayoristas? → Oportunidades B2B
- ¿Dónde se venden productos de lujo? → Expansión premium
- ¿Qué segmento genera más ingresos? → Priorización estratégica

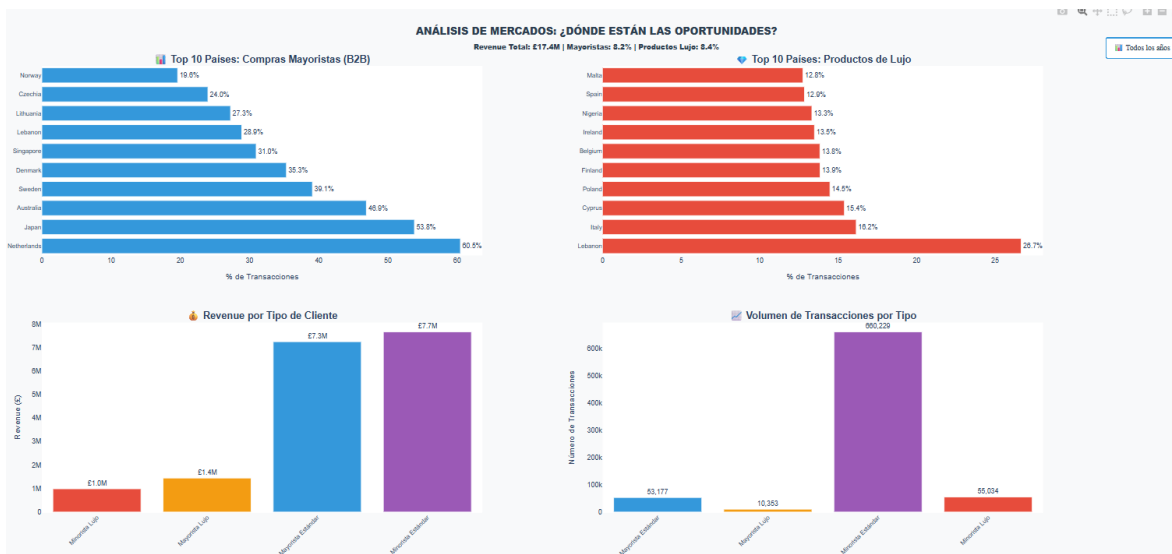


Figura 54: Dashboard Ejecutivo - Análisis de Mercados: ¿Dónde están las oportunidades?

Referencias

- [1] Michael JA Berry and Gordon S Linoff. Data mining techniques: for marketing, sales, and customer relationship management. *John Wiley & Sons*, 2004.
- [2] Dayi Chen, Sai Leung Sain, and Kun Guo. Data mining for the online retail industry: A case study of rfm model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19(3):197–208, 2012.
- [3] Peter S Fader, Bruce GS Hardie, and Ka Lok Lee. Rfm and clv: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, 42(4):415–430, 2005.
- [4] Jiawei Han, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. *Elsevier*, 2011.
- [5] Rahul Khandelwal. Customer segmentation in online retail: A detailed step-by-step explanation on performing customer segmentation in online retail dataset using python, focussing on cohort. Towards Data Science, January 1 2021. URL <https://towardsdatascience.com/customer-segmentation-in-online-retail-1fc707a6f9e6>. Accessed: 2025-09-29.
- [6] Eric WT Ngai, Li Xiu, and Dorothy CK Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2):2592–2602, 2009.

- [7] Konstantinos K Tsipitsis and Antonios Chorianopoulos. Data mining techniques in crm: inside customer segmentation. *John Wiley & Sons*, 2009.
- [8] Peter C Verhoef, Werner J Reinartz, and Manfred Krafft. Customer engagement as a new perspective in customer management. *Journal of Service Research*, 13(3):247–252, 2010.