Online Retail Data Wrangling

Índice

1. Acceso al Google Colab

El cuaderno ***OnlineRetail-DataWrangling*** se encuentra disponible en:

O por medio de un enlace directo: `Abrir en Google Colab`

2. Analizando el comportamiento de los datos

2.1. ¿Podemos describir qué es un registro?

En el dataset ***Online Retail***, 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*** 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.

2.2.1. ¿Son pocos o demasiados registros?

La cantidad de registros no es reducida. De hecho, se considera un dataset de tamaño grande, suficiente para realizar análisis exploratorios, segmentación de clientes y estudios de comportamiento de compra.

2.2.2. ¿Tenemos capacidad (CPU + RAM) suficiente para procesar?

Con **1 067 371 registros y 8 atributos**, el dataset puede manejarse sin dificultad en equipos personales con especificaciones modernas (por ejemplo, 8 GB de RAM en adelante). No obstante, si se procesan múltiples transformaciones complejas o modelos avanzados de *machine learning*, podría ser recomendable optimizar el uso de memoria o emplear técnicas de muestreo.

2.2.3. ¿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 1, mientras que la Figura 1 ilustra la distribución de estos registros en las distintas columnas del dataset.

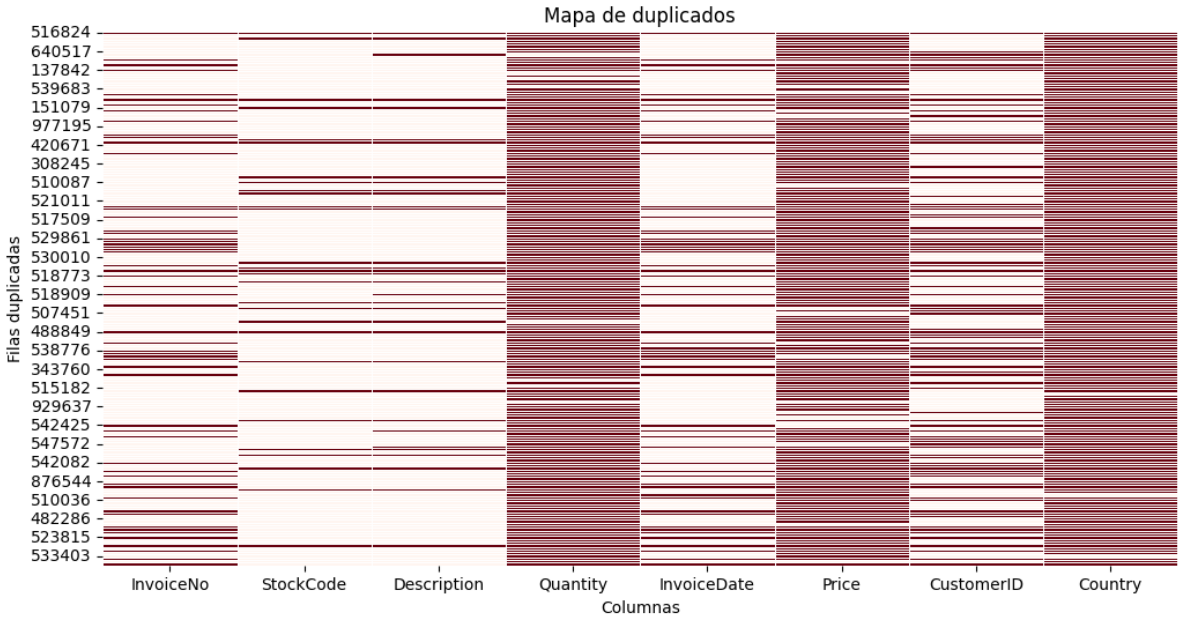


Figura 1: Mapa de calor de los registros duplicados en el dataset *Online Retail*.

Dataset	Cantidad de filas duplicadas
Online Retail	34,335

Tabla 1: Cantidad de registros duplicados detectados en el dataset *Online Retail*.

Posteriormente, se eliminaron dichos registros duplicados.

```
mostrarDuplicados(onlineRetail)
```

```
Cantidad de filas duplicadas: 0
```

```
mapaDuplicadosTodas(onlineRetail)
```

```
No hay duplicados en el DataFrame.
```

Figura 2: Resultados de la limpieza de duplicados en el dataset *Online Retail*.

2.2.4. ¿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.

2.2.5. ¿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 3 y la figura 3 resumen los resultados obtenidos.

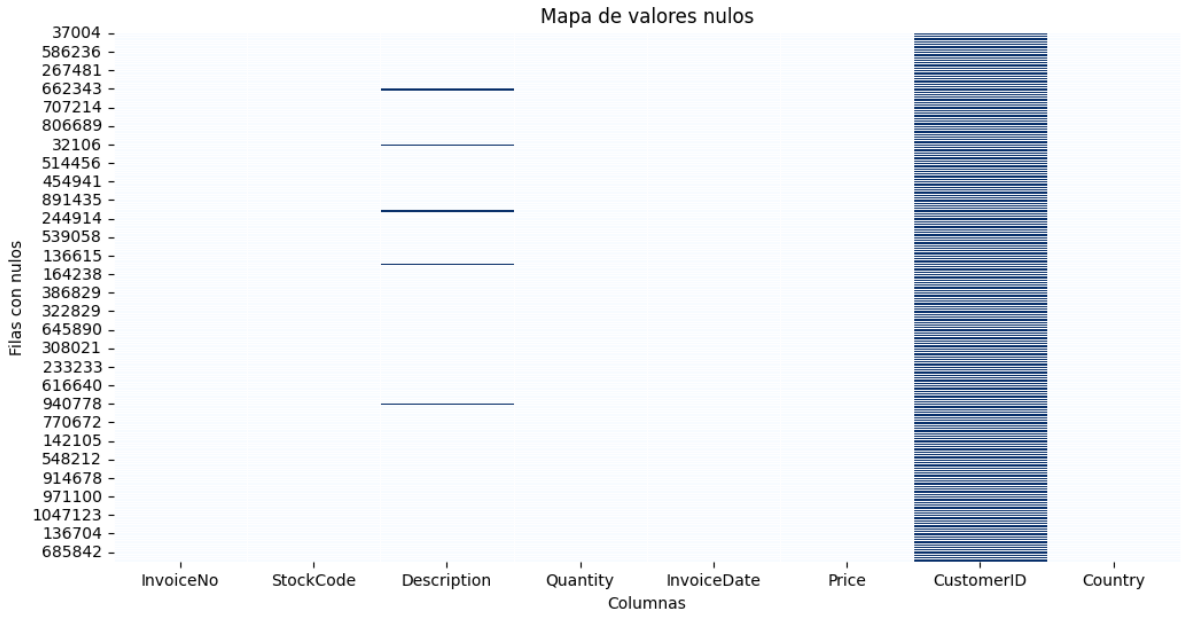


Figura 3: Mapa de calor de los valores nulos en el dataset *Online Retail* antes de la limpieza.

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

Tabla 3: Número de valores nulos por columna en el dataset *Online Retail*.

2.2.6. 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.

2.2.7. Acciones frente a los datos nulos

Hipótesis sobre valores faltantes en CustomerID En este apartado se plantea la siguiente hipótesis:

> **Hipótesis:** Los registros con valores nulos en `CustomerID` corresponden a casos en los que dicho identificador se encuentra concatenado dentro de la columna `Description`.

El interés de esta hipótesis radica en que, de ser cierta, permitiría salvar una gran cantidad de registros. En otras palabras, los valores nulos en `CustomerID` podrían ser imputados automáticamente tomando como referencia el número incrustado en la `Description`, recuperando así información de clientes que de otro modo se perdería.

En la Figura 4 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.

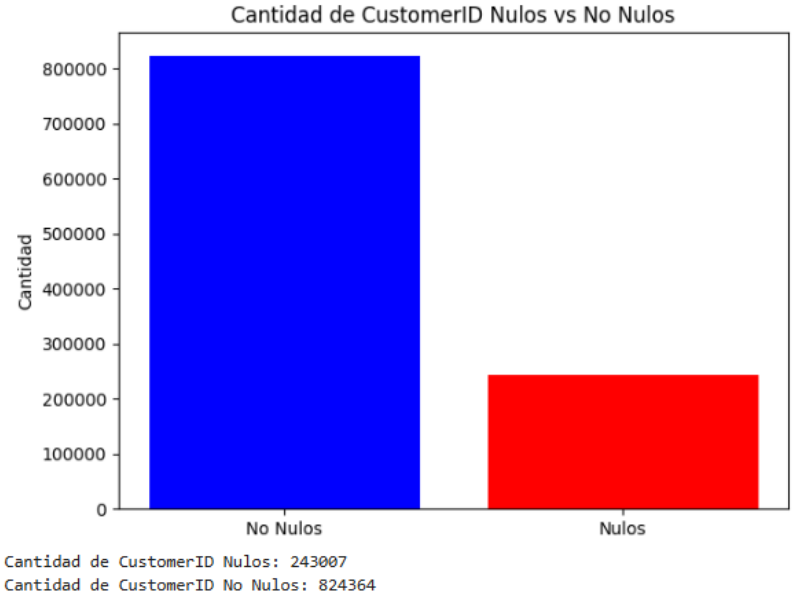


Figura 4: Comparación entre registros con `CustomerID` nulo y no nulo.

Posteriormente, se analizó la distribución del número de dígitos presentes en los identificadores válidos. La Figura 5 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`.

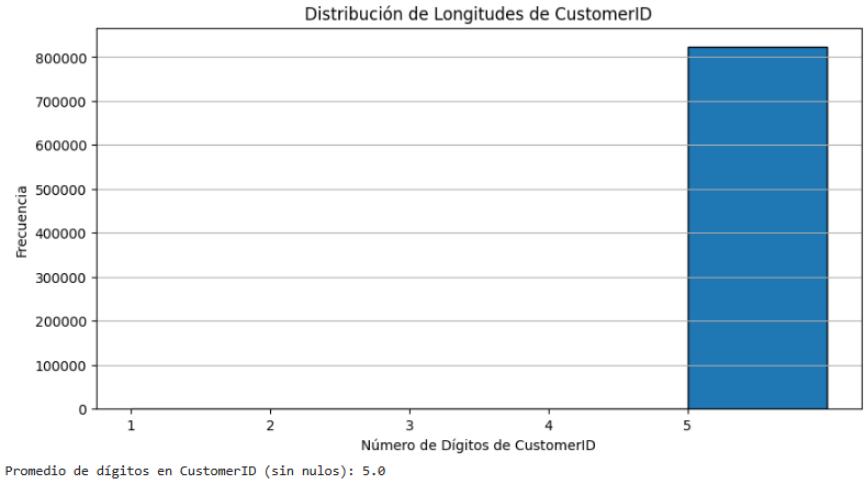


Figura 5: Distribución promedio de dígitos en los `CustomerID`.

El análisis cuantitativo arrojó los siguientes resultados:

- Cantidad de `CustomerID` nulos: 243,007.
- Cantidad de descripciones con un código de 5 dígitos: 248.

La Tabla 4 muestra que prácticamente todos los registros con un patrón de cinco dígitos en la descripción tenían `CustomerID` nulo, pero existieron dos casos donde sí había un identificador válido.

Caso	Cantidad
Descripciones con 5 dígitos y `CustomerID` no nulo	2
Descripciones con 5 dígitos y `CustomerID` nulo	246

Tabla 4: Comparación entre descripciones con 5 dígitos y valores nulos en `CustomerID`.

Finalmente, al aplicar la función `buscarFilaPorDescripcion` sobre el producto `SET 10 CARDS HANGING BAUBLES 17080`, se observó (Tabla 5) que este producto aparecía tanto con un `CustomerID` válido como con uno nulo. Esto demuestra que el número en la descripción no está ligado al identificador del cliente.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	Price	CustomerID	Country
C567690	23630	SET 10 CARDS HANGING BAUBLES 17080	-1	2011-09-21 17:01:00	2.99	15810.0	United Kingdom
567702	23630	SET 10 CARDS HANGING BAUBLES 17080	1	2011-09-22 09:38:00	2.99	15810.0	United Kingdom
579777	23630	SET 10 CARDS HANGING BAUBLES 17080	1	2011-11-30 15:13:00	2.49	NaN	United Kingdom

Tabla 5: Ejemplo de búsqueda con `buscarFilaPorDescripcion` mostrando casos con y sin `CustomerID`.

Conclusión: La hipótesis se **rechaza completamente**. Aunque algunos productos incluyen cadenas numéricas de cinco dígitos en la descripción, estas no corresponden al `CustomerID`. Por lo tanto, no es posible recuperar o imputar los valores faltantes de manera fiable a partir de la información en la columna `Description`.

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.

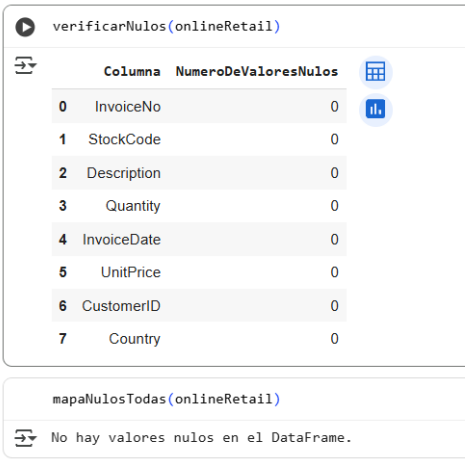


Figura 6: Salida de verificación indicando que no hay valores nulos en el dataset *Online Retail* tras la limpieza.

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

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

- `InvoiceNo`: en ciertos casos contiene letras, como la "C", que identifica facturas canceladas.

  Como se muestra en la Figura 7, se puede visualizar la proporción de facturas canceladas frente a las que no lo están.

Figura 7: Número de facturas canceladas ("C") frente a las facturas no canceladas en el dataset *Online Retail*.

La Figura 8 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 8: Salida de facturas canceladas limpias. *Online Retail*.

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

La Figura 9 muestra la distribución de las cantidades, donde se evidencian dichos valores negativos y atípicos.

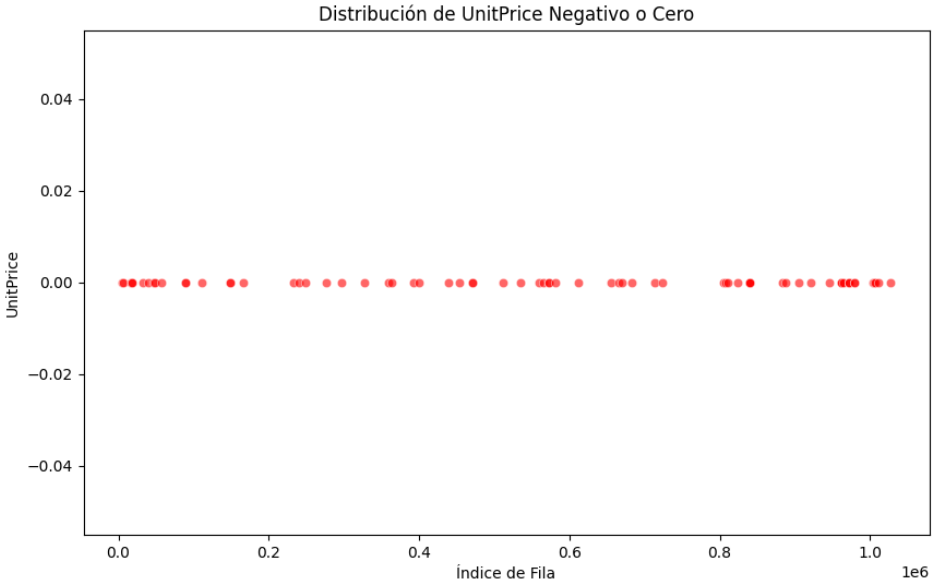


Figura 9: Distribución de la cantidad precio unitario inválidos en el dataset *Online Retail*.

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 10 muestra la nueva distribución de las cantidades y precios unitarios después del filtrado.

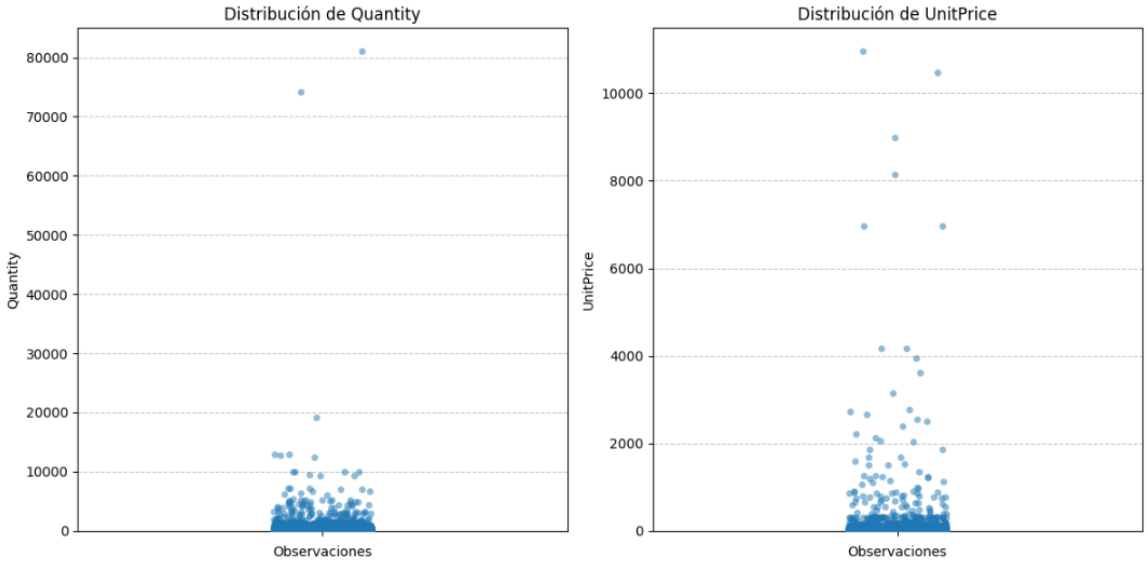


Figura 10: Distribución de la cantidad de productos por transacción y del precio unitario después del filtrado en el dataset *Online Retail*.

Adicionalmente, en el atributo `Country` se detectaron valores con codificación no estándar, como se muestra en la Tabla 6.

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 6: 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 11.

Figura 11: Proceso de normalización en `Country`.

2.2.9. ¿Entre qué rangos están los datos?

En esta sección se analizan los valores mínimos y máximos de las variables del dataset *Online Retail*, 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.
- `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.

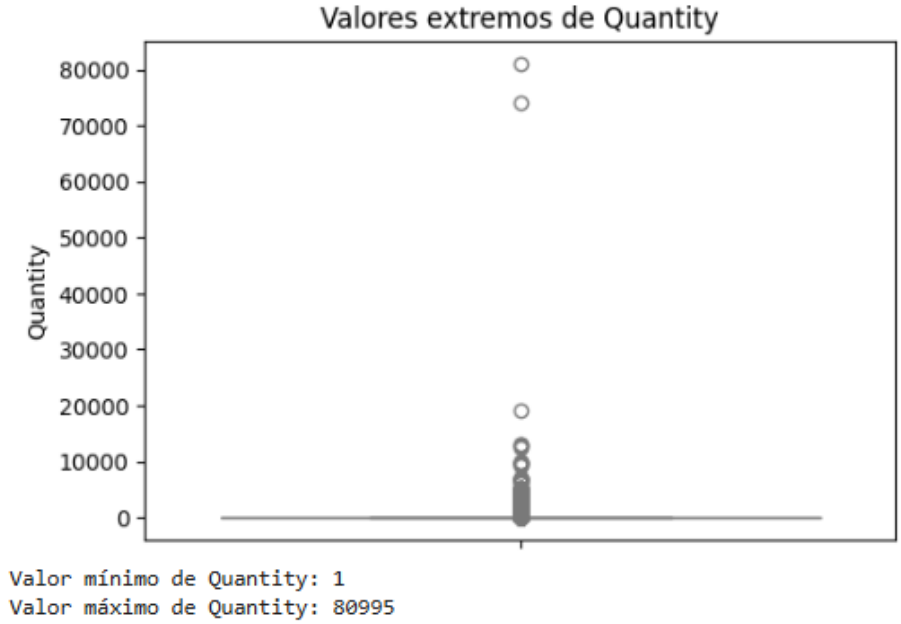
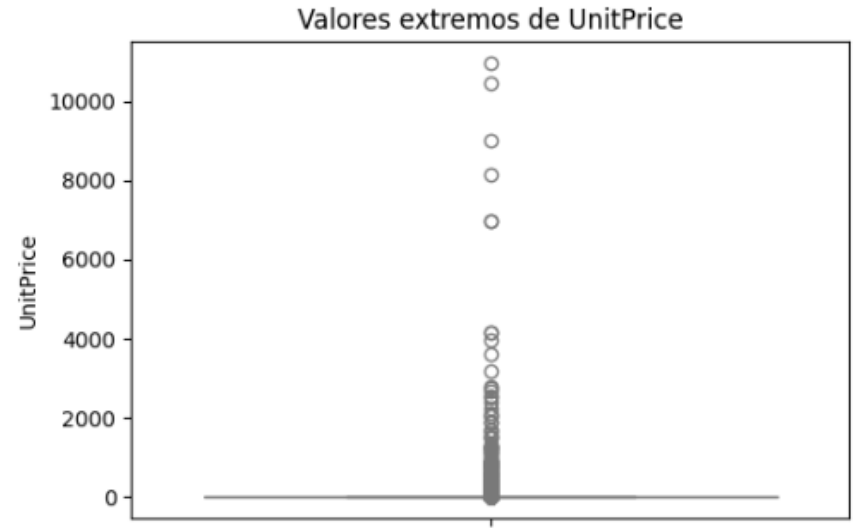


Figura 12: Distribución de la variable `Quantity`.

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.

- `UnitPrice`: rango de precios unitarios de los productos.

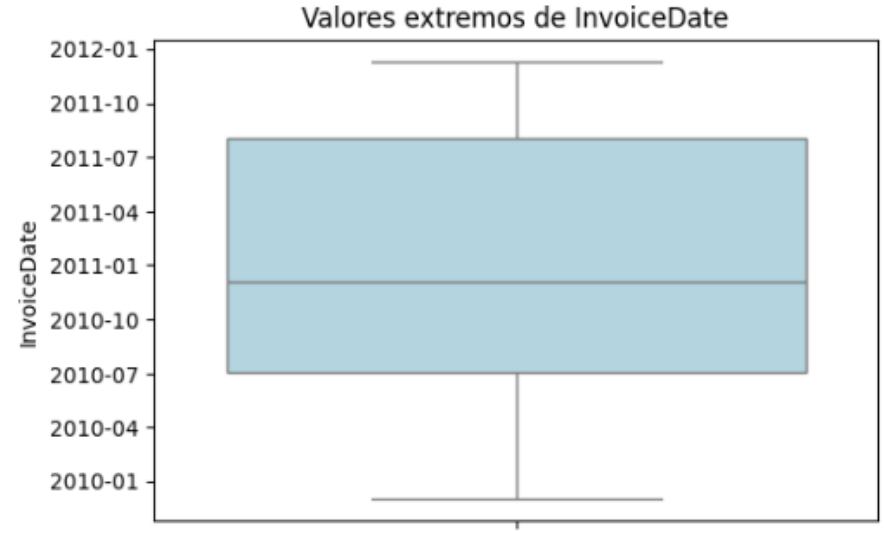



```
Valor mínimo de UnitPrice: 0.001
Valor máximo de UnitPrice: 10953.5
```

Figura 13: 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.

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




```
Valor mínimo de InvoiceDate: 2009-12-01 07:45:00
Valor máximo de InvoiceDate: 2011-12-09 12:50:00
```

Figura 14: 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.

- `CustomerID`: identificador numérico de clientes.

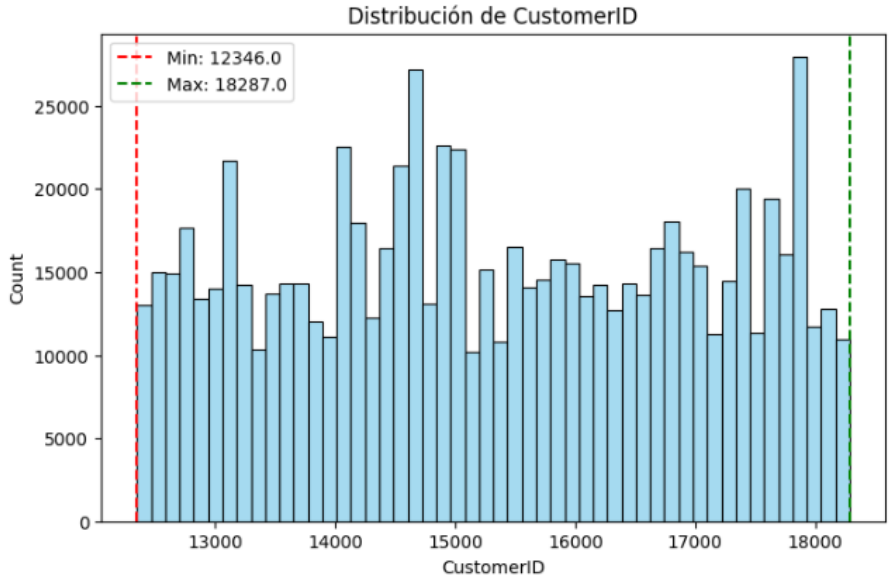


Figura 15: Distribución de la variable `CustomerID`.

El rango observado es consistente con un sistema de asignación de identificadores consecutivos.

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

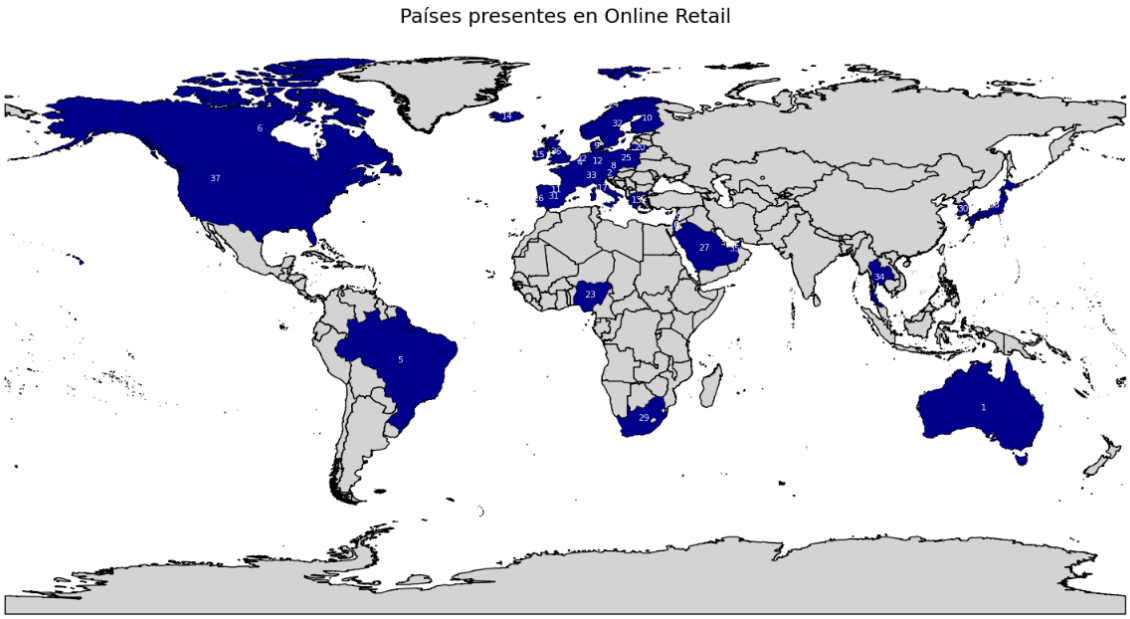


Figura 16: Mapa de países con transacciones registradas en el dataset *Online Retail*.

2.2.10. ¿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.

2.2.11. ¿Cuáles son los datos categóricos y hay necesidad de convertirlos en numéricos?

Los atributos categóricos son:

- `InvoiceNo`, `StockCode`, `Description`, `Country`.

Para algunos análisis estadísticos o modelos de aprendizaje automático será necesario convertir estas variables en numéricas mediante técnicas como *label encoding* o *one-hot encoding*.

2.3. ¿Qué representa un registro?

2.3.1. ¿Qué representa cada fila?

En el dataset ***Online Retail***, 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 7 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 7: 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 17, donde, un cliente tiene un país, un cliente tiene una a muchas detalles de facturas y un datalle de factura tiene un producto).

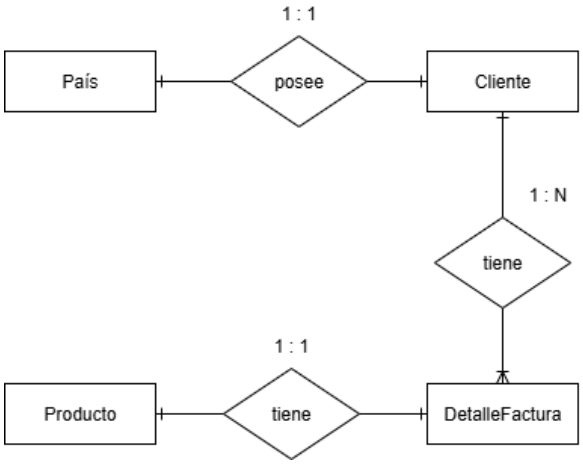


Figura 17: Diagrama entidad relación en *Online Retail*.

2.3.2. Si 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`).

2.3.3. ¿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).

2.4. ¿Siguen alguna distribución?

Para analizar la distribución de los datos en el dataset ***Online Retail***, se generó un resumen estadístico de las principales variables numéricas, cuyos resultados se presentan en la Tabla 8.

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 8: Resumen estadístico actualizado de las variables `Quantity` y `UnitPrice` en el dataset Online Retail.

A continuación se presentan interpretaciones obtenidas exclusivamente a partir de los valores de la Tabla 8.

- **Quantity:**
  - La media es 13.49 y la mediana 6; como media > 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 media > 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.

3. Análisis de outliers

3.1. ¿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 18 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.

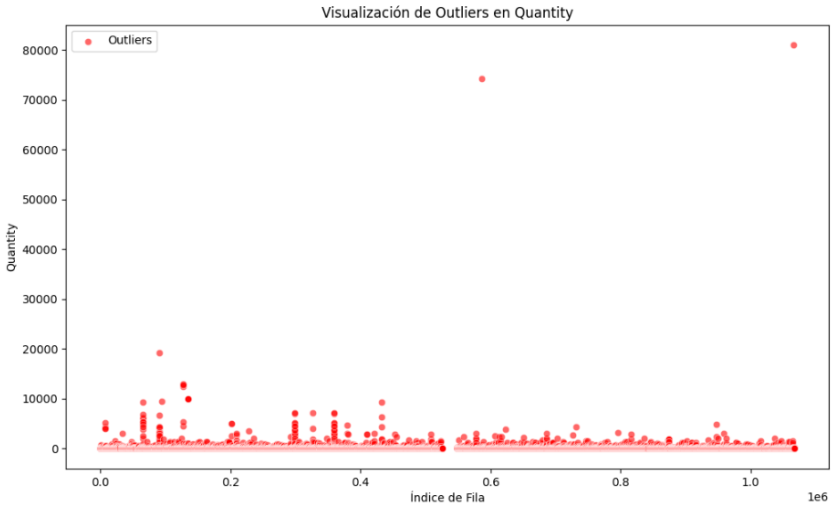


Figura 18: Visualización de outliers en `Quantity`. Los puntos rojos representan valores atípicos identificados mediante el rango intercuartílico.

De manera similar, la Figura 19 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.

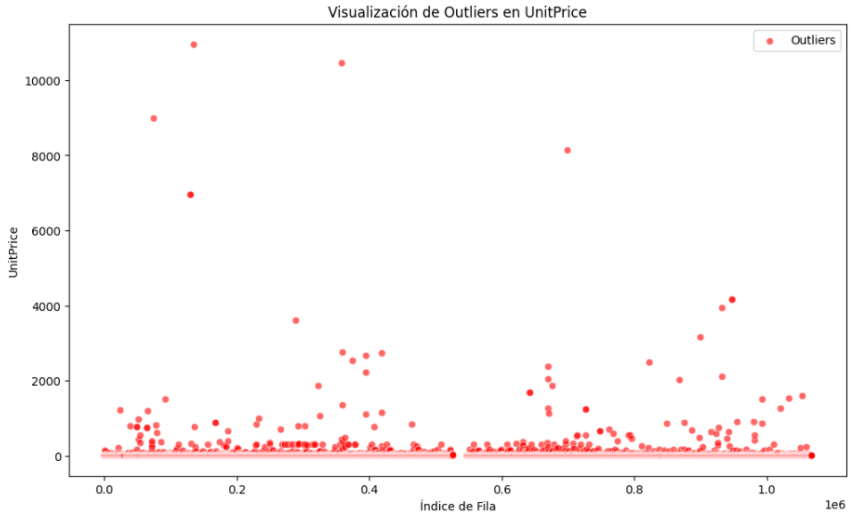


Figura 19: Visualización de outliers en `UnitPrice`. Los puntos rojos representan valores atípicos identificados mediante el rango intercuartílico.

Como referencia, se presentan también los gráficos normales sin resaltar outliers. La Figura 20 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.

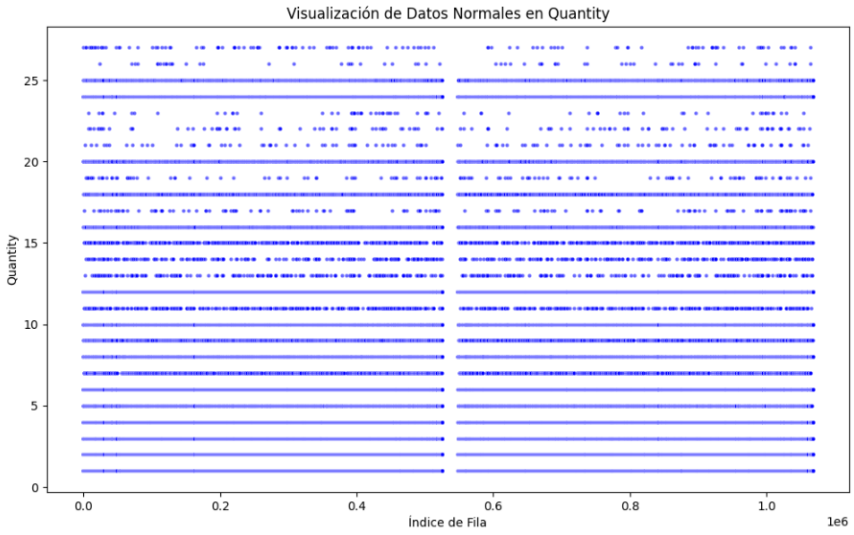


Figura 20: Distribución general de `Quantity` sin resaltar outliers. Se observa la concentración principal de transacciones en cantidades bajas.

De igual forma, la Figura 21 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.

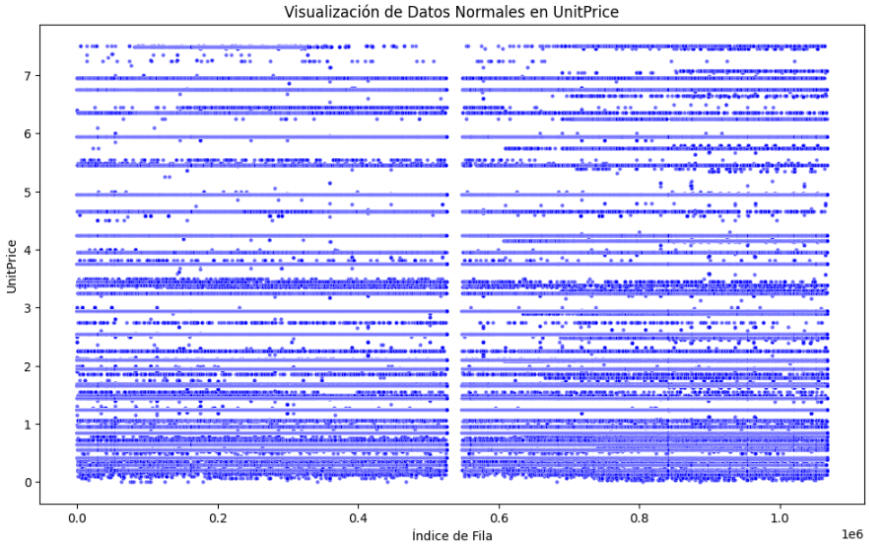


Figura 21: Distribución general de `UnitPrice` sin resaltar outliers. La mayoría de precios se encuentran dentro de un rango típico, destacando los valores extremos previamente identificados.

3.1.1. ¿Podemos eliminarlos? ¿Es importante conservarlos?

Como señala Khandelwal [1], 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 [1] 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.

Referencias

[1] 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.