OBJETIVO DEL PROYECTO

El objetivo principal de este proyecto es responder a la solicitud del cliente, una tienda especializada en ropa deportiva, para realizar un análisis de los productos vendidos y del perfil de los clientes que los adquieren. Esto implica estudiar el comportamiento de las ventas, evaluar las características de los productos ofrecidos, y segmentar la base de clientes según variables demográficas y de compra.

El análisis permitirá identificar patrones de consumo, determinar los productos más exitosos y los que requieren mayor atención, evaluar la efectividad de las estrategias y entender mejor las preferencias y comportamientos de los clientes. El propósito final es proporcionar información valiosa que apoye la toma de decisiones, optimizando así el rendimiento global de la tienda.

RAW DATA

El cliente ha proporcionado dos conjuntos de datos que constituyen la base para este análisis:

- Dataset de ventas (Sales_2023_2024.csv)
- Dataset de catálogo de productos (Product_catalog.xlsx)

Ambos archivos están vinculados por la variable común **Product_ID**, lo que facilita la integración entre la información transaccional y la descripción detallada de los productos vendidos.

Dataset de ventas (Sales_2023_2024.csv)

Contiene los registros de transacciones comerciales realizadas en la tienda, con información detallada sobre cada venta y el perfil básico del cliente. Las columnas incluyen:

- **Sale_ID**: Identificador único de cada registro de venta.
- **Product_ID**: Código del producto vendido, clave que permite unir con el catálogo de productos.
- **Store_Country**: País donde se llevó a cabo la operación de venta.
- **Quantity_Sold**: Cantidad de unidades vendidas en la transacción.
- **Discount_percent**: Porcentaje de descuento aplicado en la compra.
- **Sale_Date**: Fecha del evento de venta.
- **Payment_Method**: Canal o método de pago utilizado por el cliente.
- **Customer_Age**: Edad del comprador en el momento de la transacción.
- **Customer_Gender**: Género del cliente.
- **Membership**: Estado del cliente respecto a la membresía de la tienda.

Dataset de catálogo de productos (Product_catalog.xlsx)

Proporciona información descriptiva sobre cada producto que ofrece la tienda, con variables claves para análisis de inventario, márgenes y segmentación por producto. Las variables contenidas son:

- **Product_ID**: Identificador único que permite el enlace con los datos de venta.

- **Product_Name**: Denominación comercial del producto.
- **Category**: Clasificación del producto de acuerdo con los diferentes Departamentos de Venta de la empresa.
- **Sale_Price_EUR**: Precio de venta en euros para el cliente final. No incluye el descuento.
- **Cost_Price_EUR**: Precio de adquisición o coste para la empresa.
- **Stock**: Cantidad disponible para la venta.
- **Country_Origin**: País de fabricación o procedencia.
- Brand: Marca a la que pertenece el artículo.
- Color: Color principal del producto.
- **Size**: Talla o medida relevante para prendas y calzado.
- **Gender**: Segmento objetivo de género para el producto.

PROCESOS DEL PROYECTO

El proyecto incluye las siguientes etapas clave:

- Transformación y limpieza profunda de los datos.
- Análisis descriptivo y estadístico de los datos.
- Visualización de los datos para facilitar la interpretación
- Desarrollo de un dashboard operativo.
- Elaboración de un informe explicativo con conclusiones y recomendaciones.

HERRAMIENTAS PARA REALIZAR EL PROYECTO

Para la realización del Análisis Exploratorio de Datos (EDA) y el procesamiento de la información se emplearon herramientas robustas como Python, utilizando principalmente la librería Pandas, junto con el entorno de desarrollo Visual Studio Code, que facilitó la codificación y gestión del proyecto.

Por otro lado, la visualización de datos y el desarrollo del Dashboard interactivo se llevaron a cabo utilizando Power BI, plataforma que permitió crear representaciones visuales claras y dinámicas, facilitando la interpretación y comunicación efectiva de los resultados obtenidos.

REQUISITOS MÍNIMOS DEL PROYECTO

El proyecto final, alojado en un repositorio de GitHub, cumple con los siguientes requisitos:

- Disponer de dos conjuntos de datos en bruto provenientes de fuentes distintas, unidos para trabajar de forma conjunta.
- Conjunto de datos final transformado con una extensión mínima de 50,000 filas y 20 columnas
- Análisis exhaustivo del conjunto de datos final
- Dashboard operativo de los datos finales que aporte valor al análisis realizado.
- Informe del análisis realizado.

- Archivo README.md documentando la metodología y resumen del análisis.
- Buena organización y estructura de carpetas en el repositorio.

METÓDOLOGÍA

Para llevar a cabo el análisis de los datos, se trabajó con dos conjuntos de datos simulados en formato CSV que fueron entregados por el cliente. El proceso comenzó con una exploración preliminar de los datos (Exploratory Data Analysis, EDA) utilizando Python en el entorno de desarrollo Visual Studio Code, apoyado principalmente en la librería Pandas para la manipulación y análisis de datos.

Durante esta etapa inicial se examinó la estructura de los archivos, validando la calidad de los datos. Se confirmó que no existían registros duplicados ni valores nulos que requirieran tratamiento especial. También se realizaron las adecuaciones necesarias en el formato de las columnas: se estandarizaron los valores numéricos, reemplazando comas por puntos para asegurar su correcta interpretación como datos de tipo flotante (float), y se convirtió la información de fechas a un formato adecuado para su posterior análisis temporal.

Posteriormente, se procedió a la integración de los dos datasets a partir de la variable común Product_ID. Antes de la unión, se efectuaron las transformaciones necesarias para garantizar la coherencia y compatibilidad entre las tablas.

Para enriquecer el análisis, se crearon seis variables adicionales derivadas:

- **Gross_Sales**: Ingreso bruto por producto y transacción. (Quantity_Sold * Sale_Price_EUR)
- **Total_Discount**: Valor total del descuento aplicado. (Quantity_Sold * Discount_percent * Sale_Price_EUR)
- **Net_Sales**: Ventas netas después de descuento. (Gross_Sales – Total_Discount)
- **Total_Cost**: Costo total correspondiente a las unidades vendidas. (Quantity_Sold * Cost_Price_EUR)
- **Total_Profit_EUR**: Ganancia total calculada por producto. (Net_Sales – Total_Cost)
- **Profit_Percentage**: Porcentaje de margen de beneficio respecto a las ventas netas. (Total_Profit_EUR / Net_Sales)

El dataset final, con estas nuevas variables, fue exportado nuevamente en formato CSV para su posterior análisis avanzado y visualización en Power BI.

En la fase analítica se aplicaron técnicas de análisis descriptivo diferenciado para variables numéricas y categóricas, con el fin de identificar patrones de consumo, comportamientos de clientes y características de ventas relevantes para el negocio.

ANÁLISIS DESCRIPTIVO Y ESTADÍSTICO

Para llevar a cabo una comprensión profunda del comportamiento de las ventas y el perfil de los clientes, se realizó un análisis descriptivo y estadístico sobre los datos integrados procedentes de los dos conjuntos principales: el dataset de ventas correspondiente al periodo 2023-2024 y el catálogo de productos.

Estructura de los Datos

El conjunto de datos de ventas contiene un total de 76.578 registros con 9 columnas que describen cada transacción, mientras que el catálogo de productos aporta 1.716 registros con 11 columnas descriptivas. Tras la unión mediante la variable común Product_ID, se obtuvo un dataset final con 76.578 filas y 20 columnas. Para facilitar el análisis y la visualización, se añadieron seis columnas derivadas: Gross_Sales, Total_Discount, Net_Sales, Total_Cost, Total_Profit_EUR y Profit_Percent, ampliando el conjunto final a un total de 26 columnas.

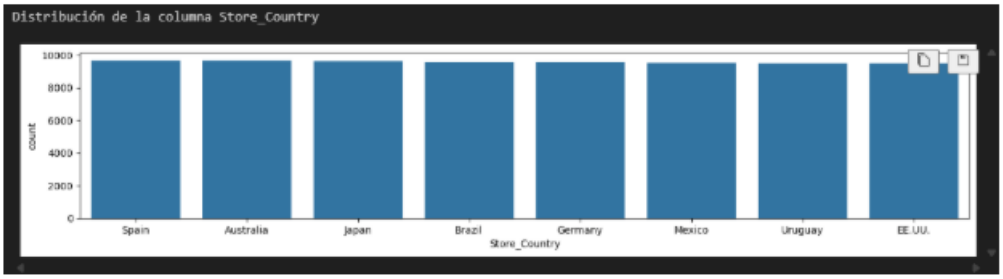
Calidad y Preparación de Datos

Se verificó que el conjunto no contenía valores nulos ni registros duplicados. Se realizó la transformación necesaria para garantizar la correcta interpretación de los datos numéricos y de fechas, estandarizando formatos y asegurando la integridad para un análisis fiable.

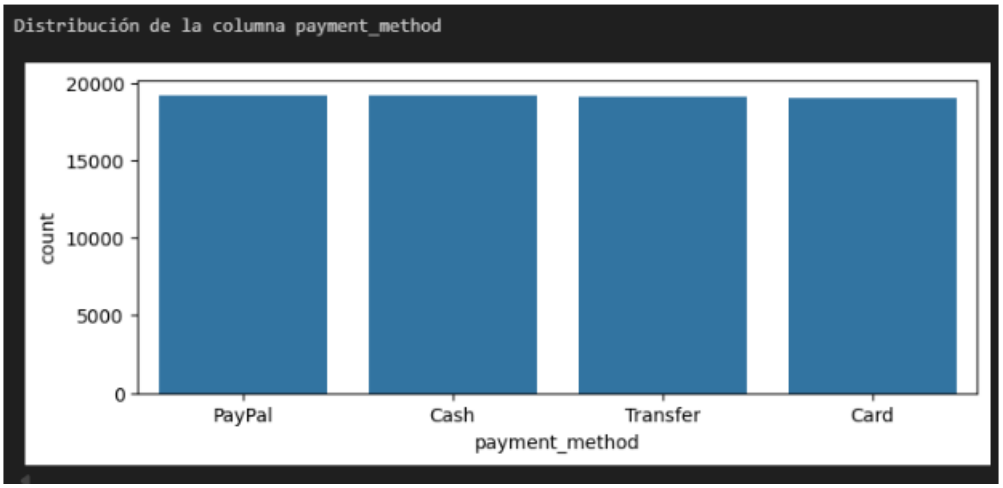
Análisis columnas categóricas

El estudio de las variables categóricas permite identificar patrones repetitivos y destacar los valores más frecuentes:

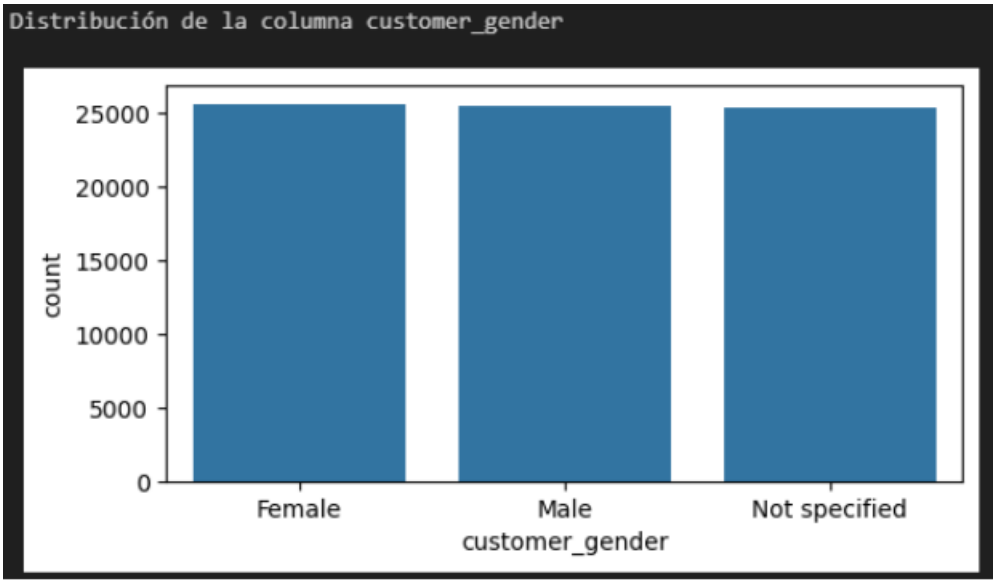
- La columna Sale_ID tiene 76.578 valores únicos
- La columna Product_ID tiene 1.716 valores únicos y el dato más repetido es S001148 con una frecuencia de 67 veces
- La columna Store_Country tiene 8 valores únicos y el dato más repetido es Spain con una frecuencia de 9.673 veces (12,6%)



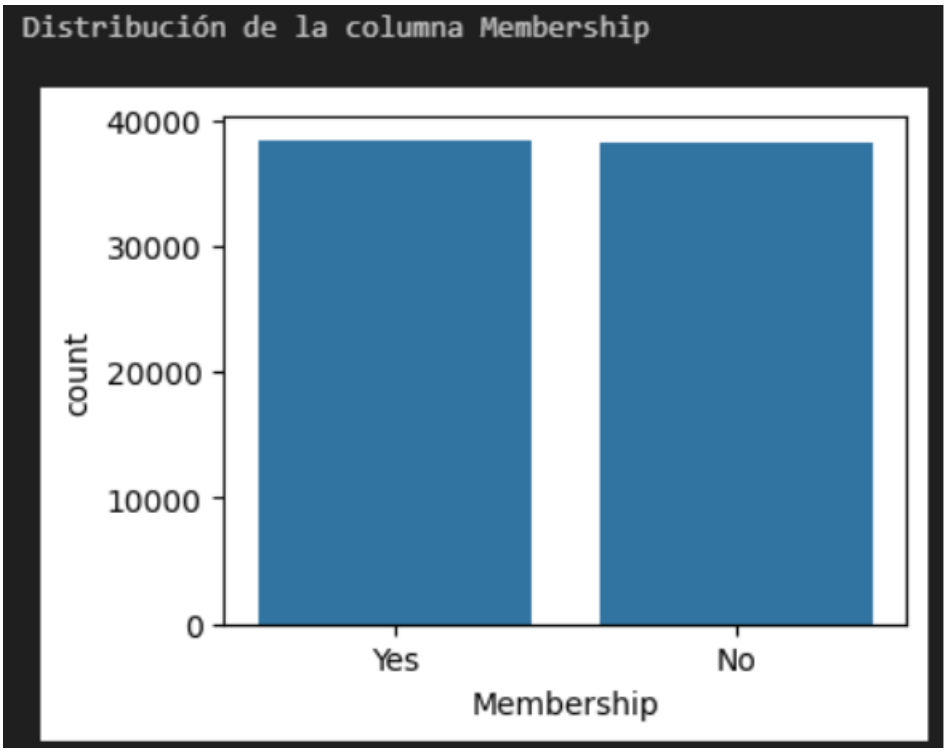
- La columna Payment_Method tiene 4 valores únicos y el dato más repetido es Paypal con una frecuencia de 19.248 veces (25,1%)



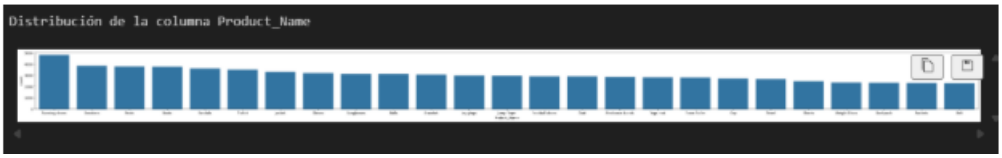
- La columna Customer_Gender tiene 3 valores únicos y el dato más repetido es female con una frecuencia de 25.669 veces (33,5%)



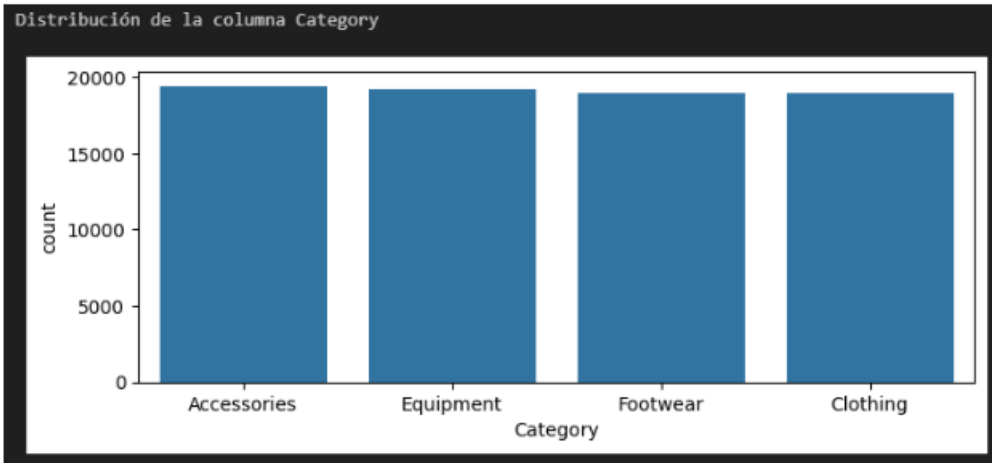
- La columna Membership tiene 2 valores únicos y el dato más repetido es Yes con una frecuencia de 38.429 veces (50,2%)



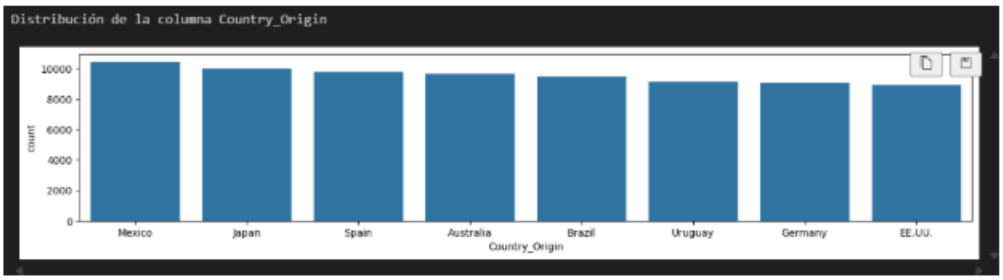
- La columna Product_Name tiene 25 valores únicos y el dato más repetido es Running shoes con una frecuencia de 4.826 veces (6,3%)



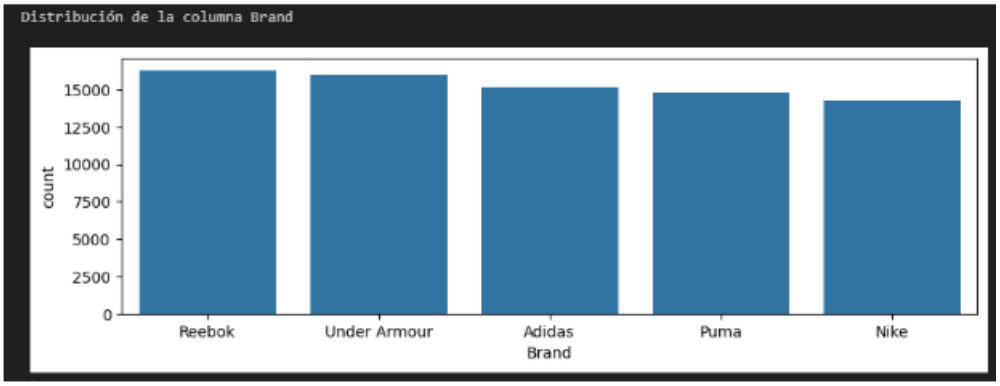
- La columna Category tiene 4 valores únicos y el dato más repetido es Accessories con una frecuencia de 19.428 veces (25,4%)



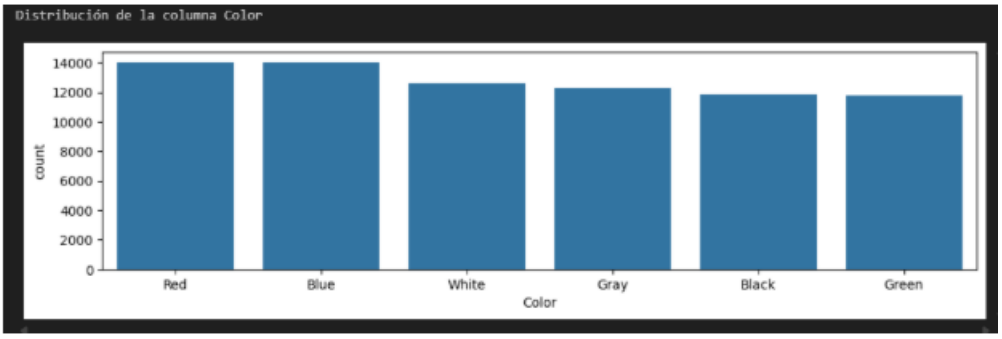
- La columna Country_Origin tiene 8 valores únicos y el dato más repetido es México con una frecuencia de 10.465 veces (13,7%)



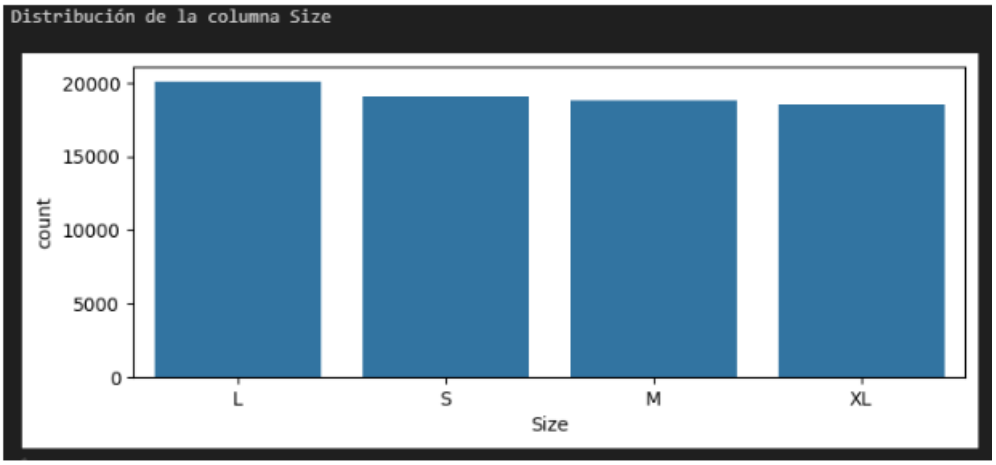
- La columna Brand tiene 5 valores únicos y el dato más repetido es Reebok con una frecuencia de 16.292 veces (21,3%)



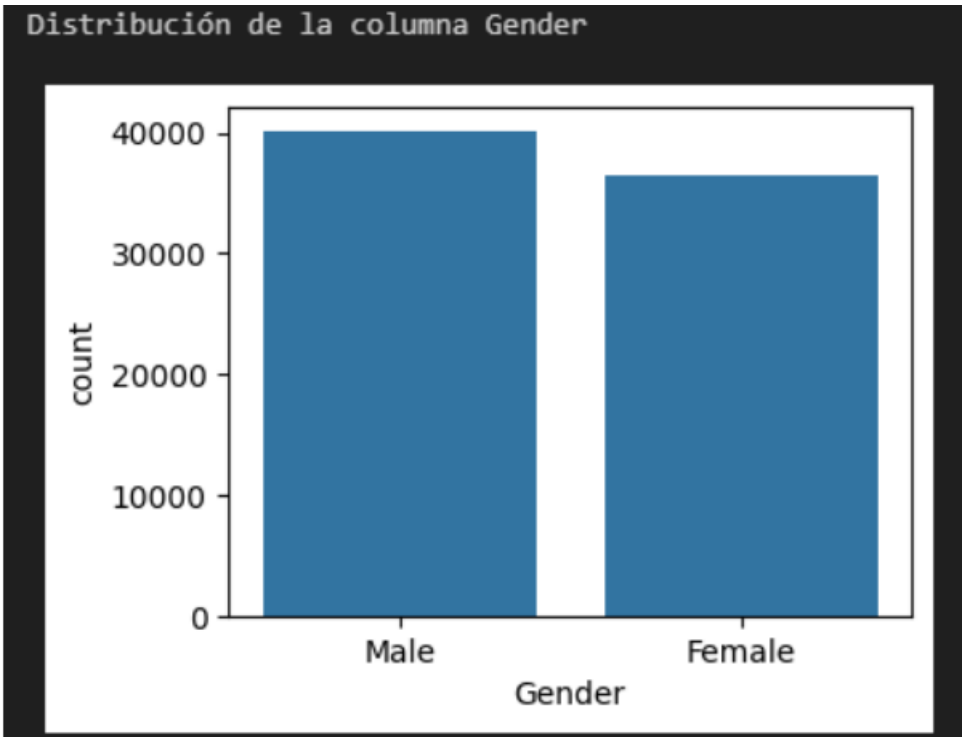
- La columna Color tiene 6 valores únicos y el dato más repetido es red con una frecuencia de 14.027 veces (18,3%)



- La columna Size tiene 4 valores únicos y el dato más repetido es L con una frecuencia de 20.095 veces (26,2%)



- La columna Gender tiene 2 valores únicos y el dato más repetido es Male con una frecuencia de 40.156 veces (52,4%)



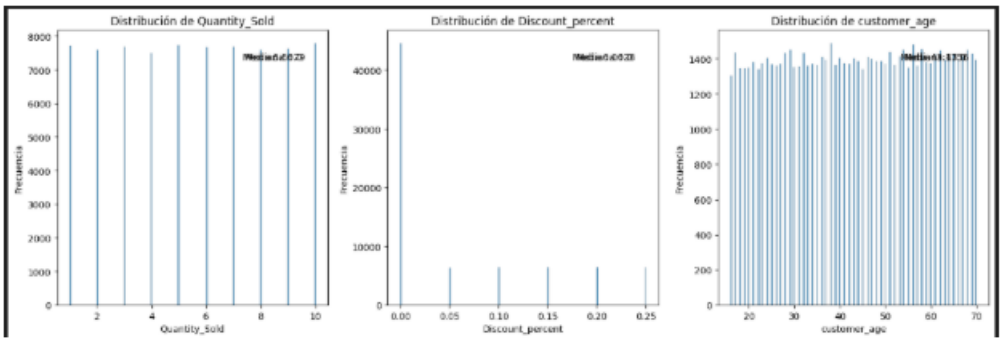
Análisis columnas numéricas

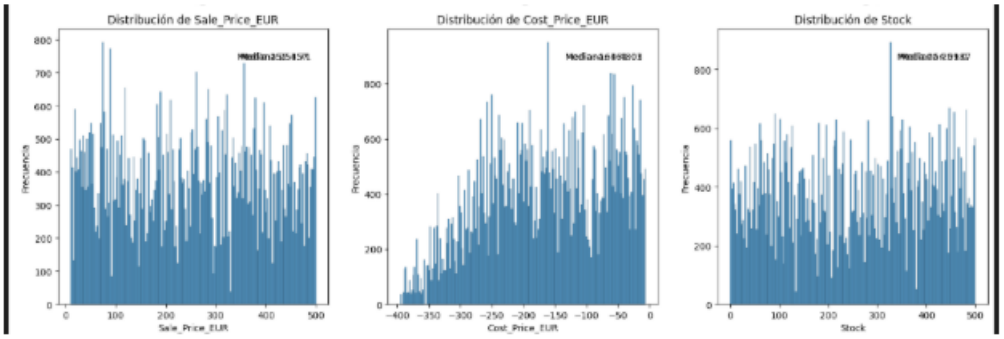
A continuación, se muestran los principales estadísticos de las columnas numéricas:

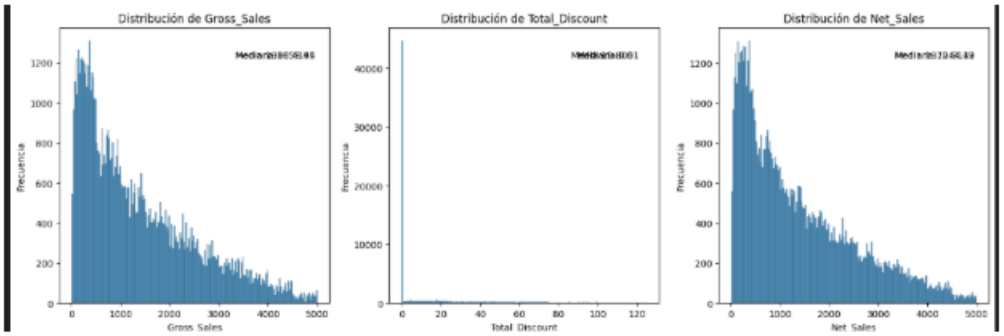
	count	mean	min	25%	50%	75%	max	std
Quantity_Sold	76578.0	5.507887	1.0	3.0	6.0	8.0	10.0	2.876896
Discount_percent	76578.0	0.062793	0.0	0.0	0.0	0.15	0.25	0.087091
sale_date	76578	2024-01-13 21:34:25.000391936	2023-01-01 00:00:00	2023-07-16 00:00:00	2024-01-26 00:00:00	2024-07-12 00:00:00	2024-12-28 00:00:00	NaN
customer_age	76578.0	43.175599	16.0	29.0	43.0	57.0	70.0	15.860606
Sale_Price_EUR	76578.0	252.156959	10.07	122.84	254.71	374.48	499.87	142.74592
Cost_Price_EUR	76578.0	-164.480334	-394.83	-241.39	-161.01	-76.51	-5.84	97.509919
Stock	76578.0	254.018674	0.0	123.0	259.0	383.0	500.0	146.33226
Gross_Sales	76578.0	1388.419357	10.07	436.545	1059.45	2089.36	4998.7	1145.986725
Total_Discount	76578.0	15.805111	0.0	0.0	0.0	22.18225	124.9675	26.72883
Net_Sales	76578.0	1372.614246	7.7475	423.1305	1044.19	2067.978375	4998.5	1140.214194
Total_Cost	76578.0	-905.701905	-3948.3	-1349.56	-680.7	-279.3	-5.84	769.363824
Total_Profit_EUR	76578.0	466.912341	-21.75	135.87	342.3165	687.698375	2428.4	417.108207
Profit_Percent	76578.0	0.337884	-0.066723	0.260013	0.339991	0.419996	0.5	0.093532

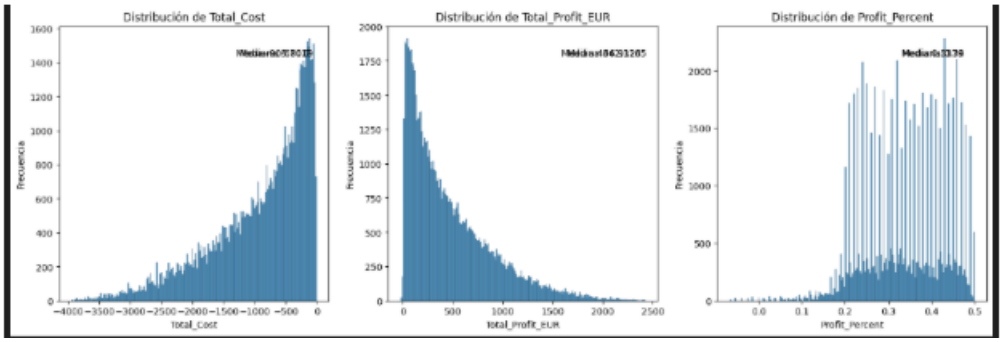
Mediante histogramas y boxplots se exploraron la distribución de las principales variables numéricas, evidenciando la variabilidad y tendencias en cantidad vendida, precios, descuentos, costos y beneficios. Esta representación gráfica sirve para identificar patrones y posibles casos atípicos.

Los histogramas son los siguientes:

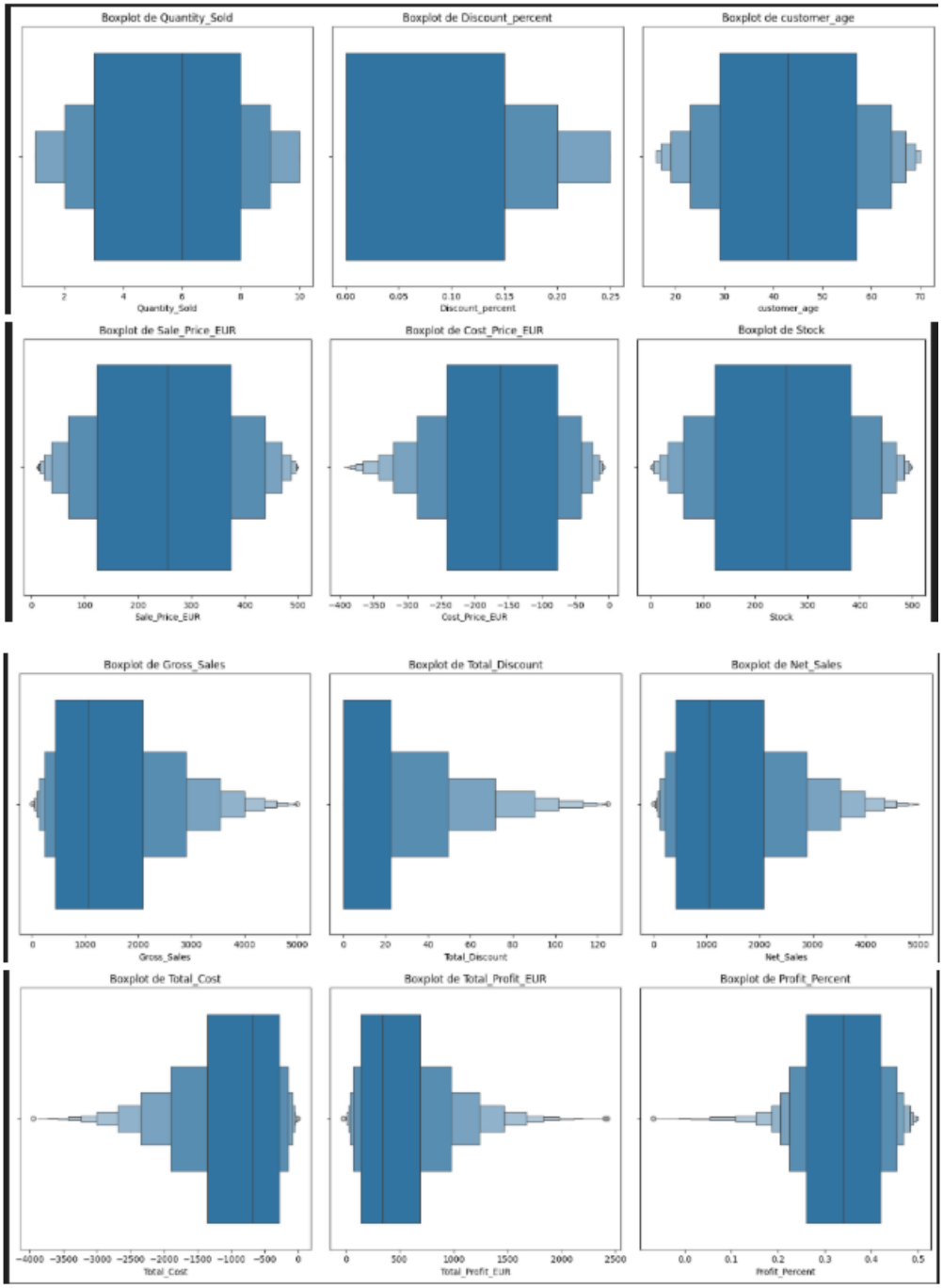








Y los boxplots son los siguientes:



No se observan outliers que puedan estar afectando las estadísticas básicas de nuestros datos.

ANÁLISIS RENTABILIDAD

Durante el periodo 2023-2024, la empresa alcanzó unas ventas netas de 99,7 millones de euros, con un margen de beneficio medio del 30,4%. En total se registraron 76.578 operaciones, con un ticket promedio de 1.301 euros por venta.

Desglose por ejercicio

- 2023: Ventas netas de 45,8 M€, margen de beneficio del 30,5% y 35.129 transacciones, con un ticket promedio de 1.303 €.
- 2024: Ventas netas de 53,9 M€, margen de beneficio del 30,4% y 41.449 transacciones, con un ticket promedio de 1.300 €.

Se observa un incremento de 8,1 M€ en ventas netas (+17,7%) en 2024 respecto a 2023. Este crecimiento se explica principalmente por el aumento en el número de operaciones (+6.320), mientras que tanto el margen como el ticket promedio permanecieron estables.

Los descuentos aplicados muestran un comportamiento distinto en cada año: en 2024 ascendieron a 3,6 M€, frente a 3,0 M€ en 2023. Además, un 33,2% de las ventas en 2024 incluyó descuentos, frente al 34,0% en 2023. Este punto requiere contraste con el cliente, ya que podría estar relacionado con factores externos de mercado o con estrategias internas de promoción.

Distribución temporal

El análisis mensual revela una evolución más sólida en 2024, con volúmenes superiores en la mayoría de meses respecto a 2023. Sin embargo, se aprecian descensos puntuales en junio, agosto, octubre y noviembre, lo que sugiere la influencia de estacionalidad o campañas específicas.

Distribución por país

La compañía opera en 8 mercados, donde las ventas se distribuyen de manera muy uniforme:

- Brasil (12,6% de las ventas netas)
- España (12,6% de las ventas netas)
- Alemania (12,5% de las ventas netas)
- Australia (12,5% de las ventas netas)
- Japón (12,5% de las ventas netas)
- México (12,5% de las ventas netas)
- Uruguay (12,4% de las ventas netas)
- Estados Unidos (12,4% de las ventas netas)

No se aprecia un país claramente dominante en términos de facturación, número de operaciones o rentabilidad.

Perfil del cliente

- Género: Las ventas se dividen de manera equilibrada entre mujeres (33,6%), hombres (33,4%) y clientes que no especificaron género (33,0%).
- Membresía: El 46,8% de las ventas corresponde a clientes con membresía activa, frente a un 53,2% de clientes no asociados.

Rendimiento por categorías de producto

La empresa agrupa su oferta en 4 departamentos, con una distribución prácticamente equilibrada:

- Clothing: 25,5% de las ventas netas
- Accessories: 25,4% de las ventas netas
- Equipment: 25,1% de las ventas netas
- Footwear: 24,0% de las ventas netas

Rendimiento por marcas

Las ventas se concentran en cinco grandes marcas deportivas internacionales, con aportaciones muy homogéneas:

- Reebok: 22,3% de las ventas netas
- Under Armour: 21,1% de las ventas netas
- Adidas: 19,6% de las ventas netas
- Puma: 18,5% de las ventas netas
- Nike: 18,5% de las ventas netas

Productos más vendidos

La empresa comercializa un total de 25 productos, destacando los siguientes en el top 5 de facturación:

- Running shoes (5,7 M€)
- Boots (5,2 M€)
- Pants (5,1 M€)
- T-shirts (4,8 M€)
- Sneakers (4,7 M€)

Es relevante señalar que, a pesar de que Footwear es el departamento con menor volumen total de ventas (24,0%) y el ticket promedio más bajo (1.260 €), concentra 3 de los 5 productos más vendidos. Esto se explica por su menor diversidad en el catálogo (5 productos) y por el ticket promedio en comparación con Clothing (6 productos / 1.343 euros ticket promedio), Accessories (7 productos / 1.303 euros ticket promedio) y Equipment (7 productos / 1.300 euros ticket promedio).