Informe de Análisis Exploratorio de Datos (EDA)

Autor: Kristhian Santiago Palomino Fajardo
Fecha: 18 de marzo de 2025

Introducción

Este informe presenta el análisis exploratorio de un conjunto de datos de ventas diarias perteneciente a una cadena minorista. El objetivo principal es comprender la distribución de las ventas, la variabilidad en el precio unitario, la estructura de las categorías de productos y el comportamiento temporal de las ventas. Este análisis es esencial para identificar patrones y tendencias que sirvan como base para el desarrollo de un modelo predictivo robusto y para la toma de decisiones estratégicas en la empresa.

Descripción del Dataset

El dataset analizado se encuentra en el archivo sales_data.csv y contiene 110 registros con las siguientes columnas:

- **Date:** Fecha de la venta, en formato string (convertida posteriormente a datetime).
- **Store:** Identificador numérico de la tienda.
- **Category:** Categoría del producto (por ejemplo, Electronics, Clothing, Home Goods).
- **Units_Sold:** Número de unidades vendidas en el día.
- **Unit_Price:** Precio unitario del producto.

Además, se han derivado nuevas variables a partir de la columna Date:

- **Mes:** Número del mes en que se realizó la venta.
- **Día_de_la_semana:** Nombre del día de la semana.
- **Trimestre:** Trimestre del año.
- **Day_of_Month:** Día del mes.
- **Es_fin_de_semana:** Indicador binario (1 para Saturday o Sunday, 0 en caso contrario).

Esta ampliación de variables temporales permite detectar patrones estacionales y comportamientos diarios que pueden influir en las ventas.

Estadísticas Descriptivas

Las estadísticas descriptivas se presentan a continuación para las dos variables numéricas clave, Units_Sold y Unit_Price:

	count	mean	std	min	25%	50%	75%	max
Units_Sold	110.0	35.309091	12.986758	15.00	25.25	32.00	44.25	62.00
Unit_Price	110.0	121.444545	125.711453	19.99	19.99	49.99	299.99	299.99

Interpretación:

- **Units_Sold:** Con un promedio de 35.31 y un rango de 15 a 62, se observa una variabilidad moderada. Los cuartiles sugieren que el 25% de las ventas son inferiores a 25.25 unidades y el 75% son inferiores a 44.25 unidades.
- **Unit_Price:** Aunque la media es de 121.44, la gran diferencia entre el 25% (19.99) y el 75% (299.99) indica la presencia de productos con precios muy distintos, lo que probablemente se relaciona con las diferentes categorías de productos.

Calidad de los Datos

Se revisó la existencia de valores nulos en el dataset y no se detectaron valores nulos en ninguna de las columnas, lo que sugiere una buena calidad de datos y facilita el análisis y modelado posterior.

Análisis de Frecuencias y Distribuciones

Frecuencia de Tiendas y Categorías

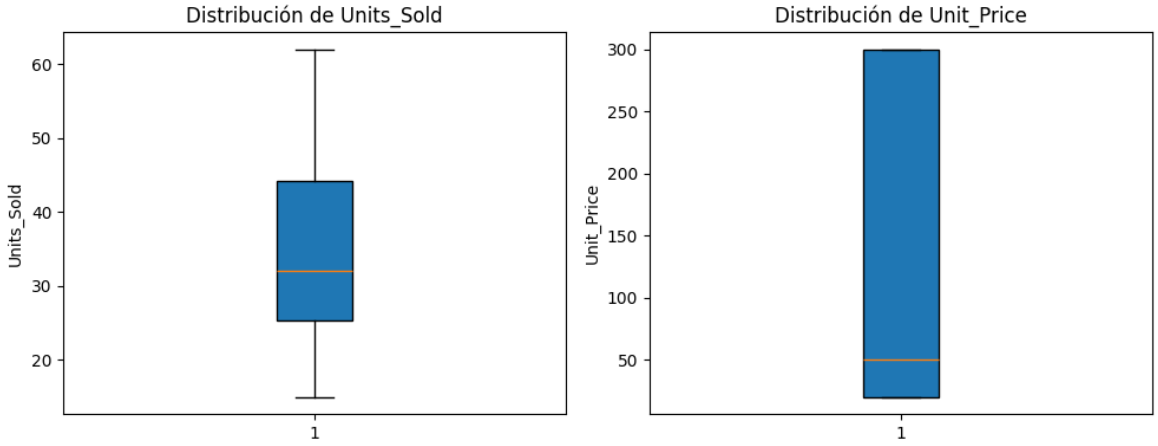
- **Tiendas:**
    - Tienda 103: 38 registros
    - Tienda 101: 36 registros
    - Tienda 102: 36 registros
- **Categorías:**
    - Home Goods: 38 registros
    - Electronics: 36 registros
    - Clothing: 36 registros

Interpretación:

El dataset está equilibrado en cuanto a la cantidad de registros por tienda y por categoría. Sin embargo, se observa que cada tienda vende únicamente una categoría específica, lo cual puede simplificar el análisis de comportamiento, pero también indica que estrategias de inventario y marketing pueden diferir sustancialmente entre tiendas.

Visualizaciones

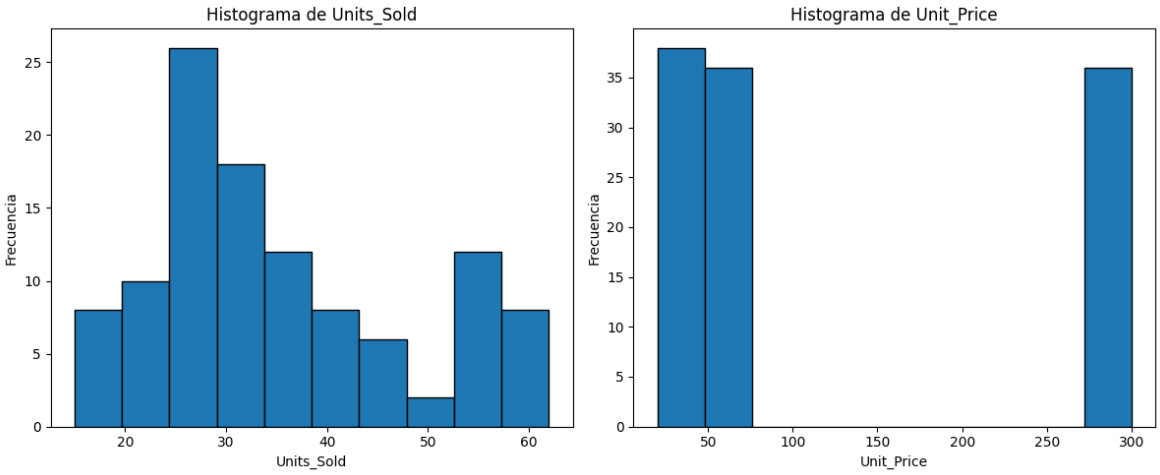
Boxplots



Los boxplots muestran la distribución de Units_Sold y Unit_Price:

- **Units_Sold:** La mayoría de los datos se agrupan en un rango intercuartílico que va aproximadamente de 25 a 44, con algunos valores atípicos menores.
- **Unit_Price:** Se evidencia una diferencia extrema en los precios, donde se observa que hay un grupo de productos a precios bajos (alrededor de 20) y otro grupo con precios altos (alrededor de 300).

Histogramas



Los histogramas confirman:

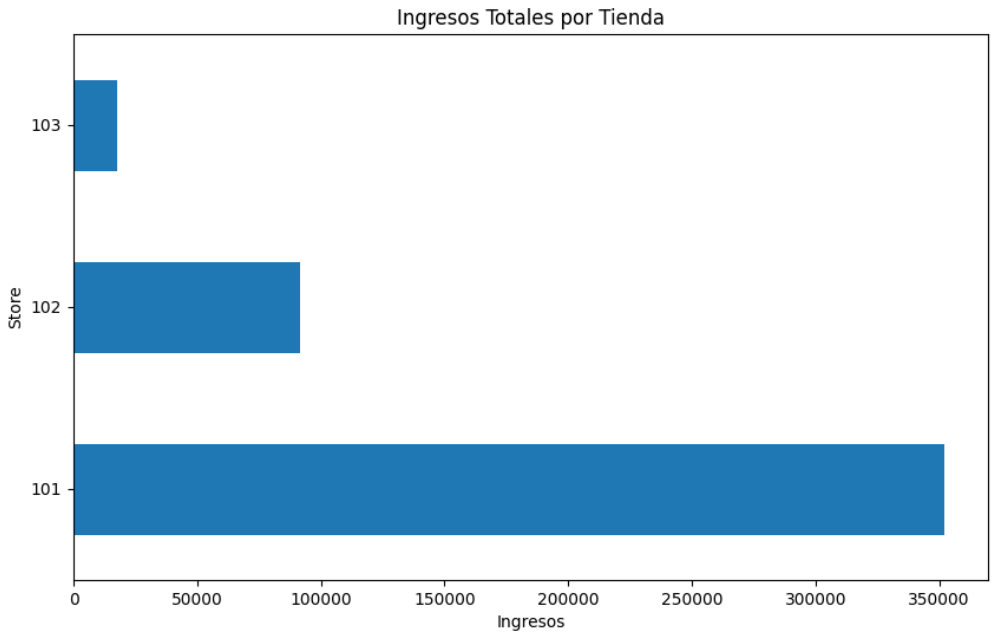
- Una distribución de Units_Sold concentrada principalmente entre 25 y 40 unidades.

- Una distribución bimodal en Unit_Price, lo cual refuerza la idea de que cada categoría tiene un rango de precios distinto (por ejemplo, Electronics a precios altos, Clothing y Home Goods a precios más bajos).

Ingresos Totales

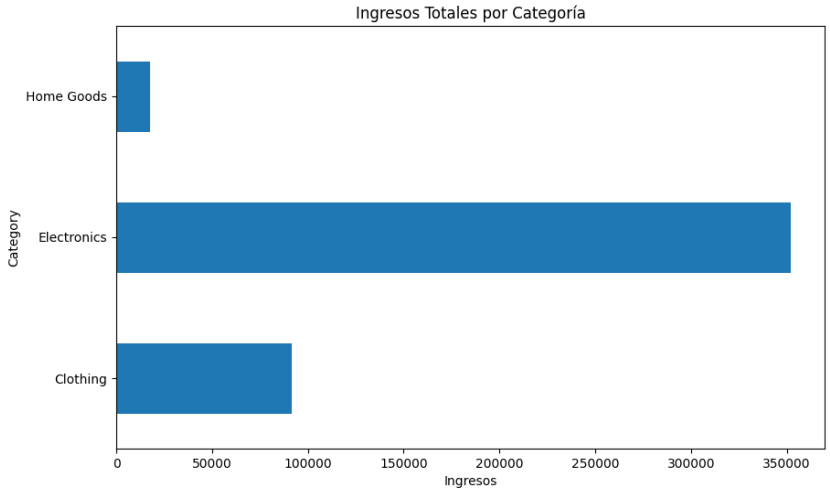
Los ingresos se calcularon como el producto de Units_Sold y Unit_Price.

- **Ingresos por Tienda:**



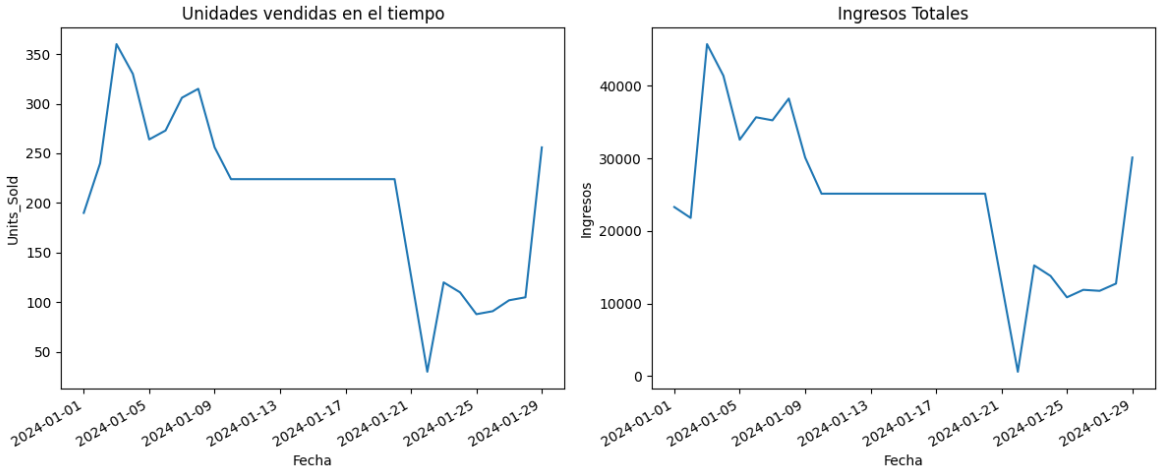
Se observa que la Tienda 101 (Electronics) registra ingresos muy elevados, lo cual se debe al alto precio unitario.

- **Ingresos por Categoría:**



La categoría Electronics domina en términos de ingresos, seguida por Clothing y Home Goods.

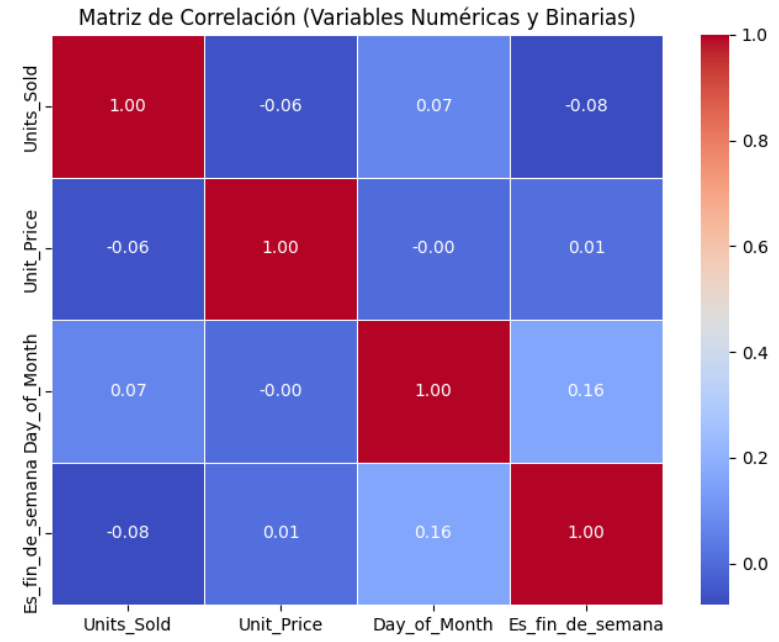
Series de Tiempo



El análisis de series de tiempo muestra:

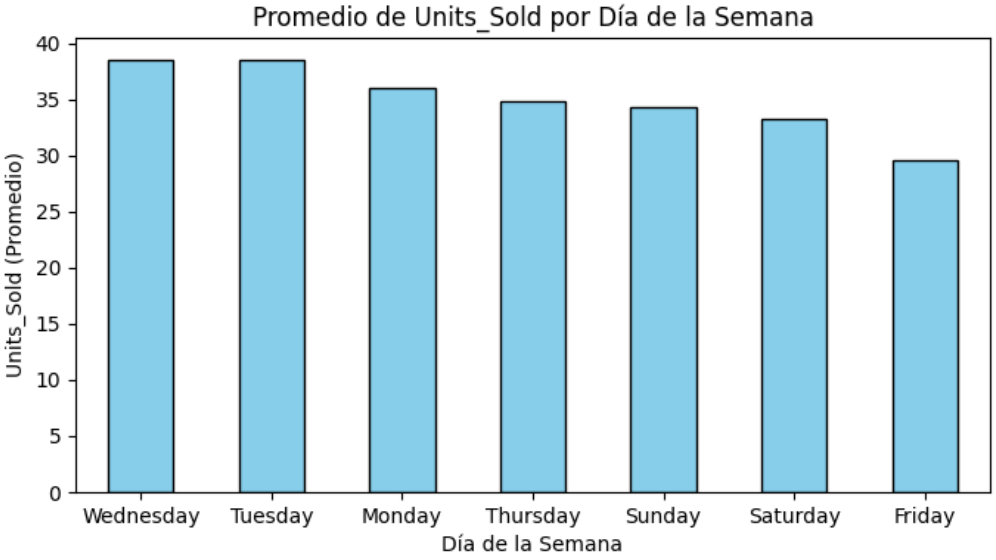
- Picos y caídas en las ventas y los ingresos a lo largo del mes.
- Una posible tendencia de disminución a mitad de mes, seguida de un repunte al final, lo que sugiere patrones estacionales o de comportamiento del consumidor.

Análisis de Correlación



La matriz de correlación revela relaciones débiles entre las variables analizadas, lo que sugiere que ninguna variable por sí sola explica fuertemente el comportamiento de las ventas (Units_Sold). Las correlaciones entre variables predictoras son mínimas, lo que valida la inclusión de todas en el modelo sin riesgo de redundancia.

Análisis por Día de la Semana

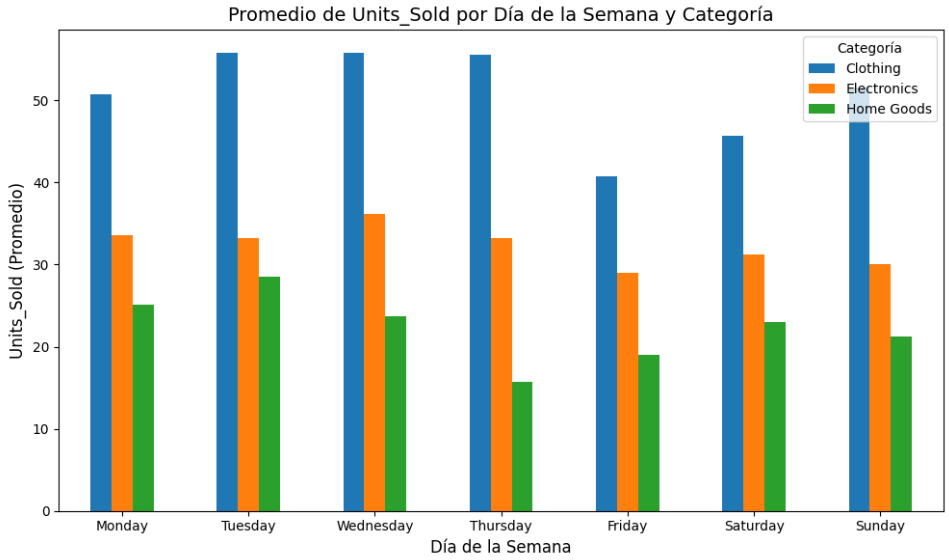


El promedio de Units_Sold por día de la semana revela:

- **Miércoles:** 38.56 unidades (máximo promedio).
- **Martes:** 38.50 unidades.
- **Viernes:** 29.58 unidades (mínimo promedio).

Esto sugiere que los días centrales de la semana (martes y miércoles) tienen mayor actividad, mientras que el viernes se comporta de manera más baja, lo que puede influir en la planificación de promociones y la asignación de recursos.

Análisis por Día de la Semana y Categoría



Al comparar las tres categorías (Clothing, Electronics y Home Goods) en cada día de la semana, se observan las siguientes tendencias:

Electronics

- Es la categoría con el promedio de ventas más alto en la mayoría de los días, superando con frecuencia las 40 unidades en los picos de actividad (por ejemplo, lunes, martes y miércoles).
- Su nivel de ventas se mantiene por encima de las demás categorías, lo que sugiere una demanda constante a lo largo de la semana.

Clothing

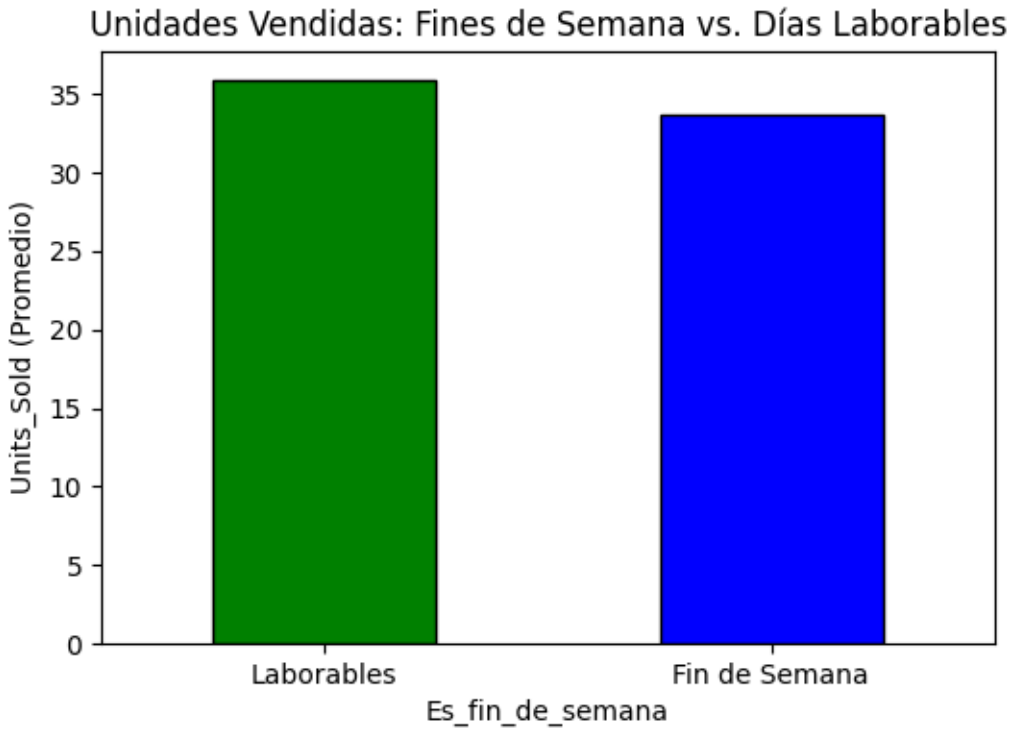
- Presenta un promedio intermedio de ventas.
- Se acerca a Electronics durante los días de mayor actividad general (martes y miércoles), con promedios que rondan las 35 unidades en algunos casos.

- Esto indica que los días centrales de la semana son también relevantes para impulsar las ventas de Clothing mediante promociones o surtido adecuado.

Home Goods

- Exhibe los valores más bajos de las tres categorías, con un promedio que puede oscilar entre 20 y 25 unidades según el día.
- Sin embargo, se aprecia un ligero repunte hacia el fin de semana (especialmente sábado), lo que sugiere una oportunidad de enfocar promociones o esfuerzos de marketing en esos días para incrementar la venta de productos para el hogar.

Comparación Fines de Semana vs. Días Laborables

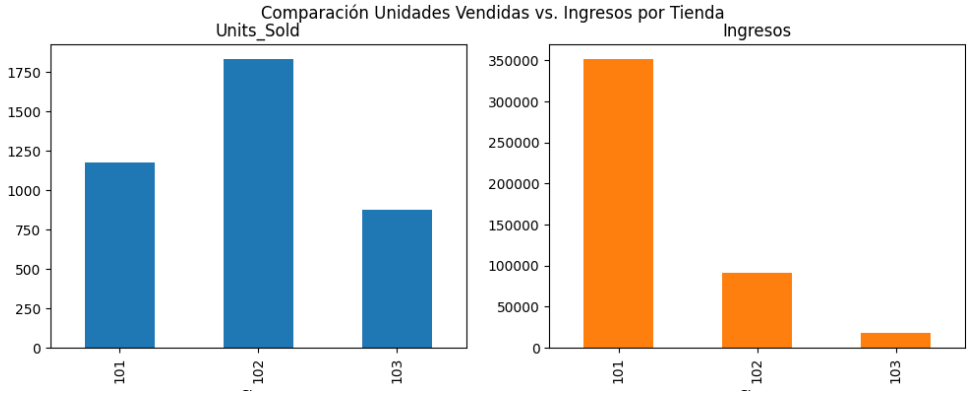


Tras clasificar los días en laborables y fines de semana (sábado y domingo), se calculó el promedio de unidades vendidas:

- Días Laborables: ~35.9 unidades en promedio.
- Fines de Semana: ~33.7 unidades en promedio.

Esta ligera disminución en fines de semana podría indicar que los clientes realizan la mayor parte de sus compras en días laborables, o que las tiendas manejan inventario/horarios distintos el fin de semana. Para el equipo de ventas, esto sugiere la posibilidad de enfocar promociones que incentiven las compras de sábado y domingo, o bien, reforzar la dotación de personal y stock en días laborables donde la demanda es consistentemente mayor.

Comparación de Unidades Vendidas vs. Ingresos por Tienda



A partir de la variable Ingresos (calculada como Units_Sold * Unit_Price), se comparó Unidades Vendidas (Units_Sold) y Ingresos Totales (Ingresos) por tienda:

1. **Tienda 101:**
    - ~1,174 unidades vendidas en total, con ingresos de ~352K.
    - Refleja un precio unitario alto (Electronics) y contribuye fuertemente a los ingresos.
2. **Tienda 102:**
    - ~1,834 unidades vendidas, ingresos de ~91K.
    - Vende más unidades que 101, pero con menor precio unitario (Clothing).
3. **Tienda 103:**
    - ~876 unidades vendidas, ~17.5K en ingresos.
    - Home Goods, con menor precio unitario y menor volumen de ventas totales.

La Tienda 101 (Electronics) domina en términos de ingresos, a pesar de no tener el mayor volumen de ventas. Esto indica la importancia de la categoría con precio elevado, que compensa un menor número de unidades vendidas. Por otro lado, la tienda 102 (Clothing) tiene la mayor cantidad de unidades vendidas, pero menos ingresos. Esto orienta la estrategia comercial:

- **Tienda 101:** Focalizar promociones de alto valor y asegurar inventario de productos electrónicos.
- **Tienda 102:** Dado que vende más unidades, se pueden diseñar ofertas de cross-selling o bundle para elevar ingresos totales.

- **Tienda 103:** Requiere mayor impulso de marketing o diversificación de productos para mejorar ingresos.

Hallazgos Clave

- **Distribución y Variabilidad:**
  Las estadísticas descriptivas indican una variabilidad moderada en Units_Sold, pero una dispersión muy amplia en Unit_Price, lo cual es consistente con la presencia de productos de muy alto y muy bajo costo.
- **Equilibrio en el Dataset:**
  Aunque el número de registros es relativamente uniforme entre tiendas y categorías, se observa que cada tienda vende únicamente una categoría, lo que sugiere que la estrategia de ventas y marketing debe adaptarse a las particularidades de cada tienda.
- **Impacto en los Ingresos:**
  La alta concentración de ingresos en Electronics (Tienda 101) destaca la importancia de gestionar el inventario y las promociones en esta categoría, a pesar de que las unidades vendidas no sean las mayores.
- **Patrones Temporales:**
  La serie de tiempo sugiere fluctuaciones a lo largo del mes, lo que podría asociarse a factores estacionales o eventos puntuales.

Conclusiones y Recomendaciones

1. **Estrategia de Inventario y Promociones:**
   - Se recomienda enfocar promociones y gestión de inventario en Electronics, dada su alta contribución a los ingresos.
   - Los días martes y miércoles, que presentan mayores ventas, podrían beneficiarse de campañas promocionales específicas.
2. **Optimización de Precios:**
   - La existencia de dos grupos de precios sugiere que la estrategia de precios debe ser diferenciada por categoría.
   - Se debe evaluar si existe potencial para ajustar precios en las categorías de bajo costo para mejorar márgenes sin afectar el volumen de ventas.
3. **Siguientes Pasos:**

- Utilizar este análisis como base para la ingeniería de características en la construcción del modelo predictivo.

Limitaciones y Consideraciones

- **Datos Limitados:**
  El análisis se basa en datos de un único mes, lo que puede limitar la detección de patrones estacionales a largo plazo.
- **Falta de Variables Externas:**
  No se dispone de información sobre promociones, eventos o factores externos que podrían influir en las ventas, lo que podría enriquecer el análisis.
- **Distribución por Tienda-Categoría:**
  La relación uno a uno entre tienda y categoría simplifica el análisis, pero también limita la posibilidad de generalizar estrategias a otras tiendas con múltiples categorías.