

PROBLEMÁTICA DE NEGOCIO

Contexto

La empresa es una tienda minorista tipo Superstore que vende productos en distintas:

- Regiones
- Categorías
- Segmentos de clientes

El negocio quiere mejorar la toma de decisiones comerciales y operativas.

Objetivo del análisis

Analizar el comportamiento de ventas de una tienda tipo Superstore para identificar patrones por categoría, región, segmento y productos extremos. Analizar el comportamiento de las ventas para:

- Identificar patrones por categoría, región y segmento
- Comprender la distribución de las ventas
- Detectar productos con desempeño extremo (ventas muy altas y muy bajas)
- Extraer conclusiones accionables desde un enfoque de negocio

El análisis se centra en ventas (sales) como métrica principal, debido a las limitaciones del dataset.

Dataset

- **Fuente:** Kaggle – Superstore Sales Dataset
- **Link:** <https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting>
- **Registros:** 9789 filas
- **Columnas principales:**
 - Order Date
 - Ship Date
 - Category
 - Sub-Category
 - Region
 - Segment
 - Product ID
 - Sales

Limitaciones del dataset

- No se dispone de información de:
 - Costos
 - Cantidad de unidades
 - Margen o profit
- Existen valores faltantes (NaT) en columnas de fechas, especialmente en *Ship Date*

Estas limitaciones se tuvieron en cuenta para evitar conclusiones incorrectas.

Proceso de limpieza de datos

El proceso de limpieza se realizó utilizando Python (Pandas) y contempló:

- Conversión correcta de columnas de fecha a formato datetime
- Tratamiento consciente de valores NaT, evitando su eliminación masiva para no perder más del 60% del dataset
- Estandarización de columnas categóricas (mayúsculas/minúsculas, espacios)
- Creación de variables derivadas:
 - Año y mes de la orden
- Validación de consistencia general del dataset
- Visualización (Matplotlib / Seaborn)

El objetivo fue obtener un dataset apto para análisis exploratorio y visualización, sin alterar la naturaleza de los datos.

Preguntas clave que el negocio necesita responder

Estas son las **preguntas reales** que justifican tu análisis:

Ventas

1. ¿Qué categorías y subcategorías generan más ingresos?
2. ¿En qué regiones se vende más?
3. ¿Qué segmentos de clientes son más valiosos?

Tendencia

4. ¿Cómo evolucionan las ventas a lo largo del tiempo?
5. ¿Existen patrones estacionales?

Logística

6. ¿Cuánto tarda, en promedio, un pedido en ser enviado?
7. ¿Existen regiones o categorías con mayores tiempos de envío?

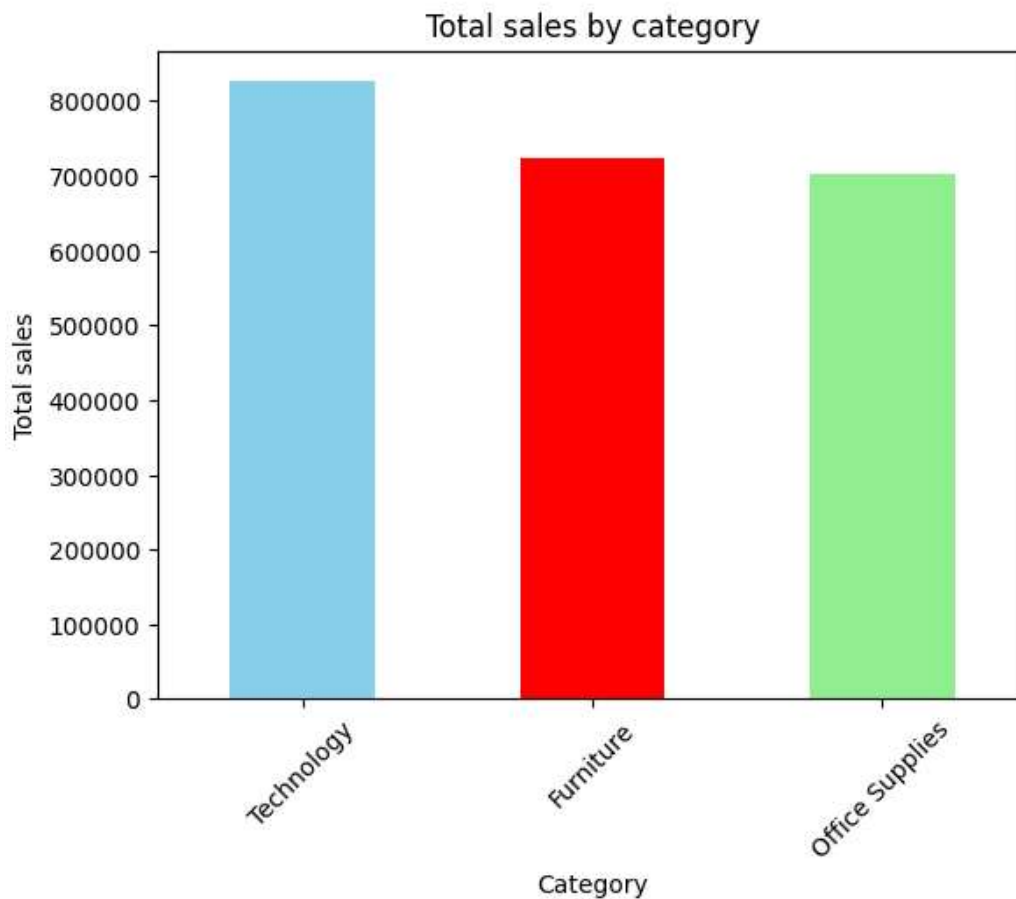
Enfoque de análisis

El análisis se desarrolló en cuatro bloques principales:

1. **Análisis general de ventas**
 - Ventas por categoría
 - Ventas por región
 - Ventas por segmento de cliente
2. **Análisis temporal**
 - Evolución de ventas en el tiempo
 - Evaluación del impacto de datos faltantes en fechas
3. **Distribución de ventas**
 - Identificación de asimetría
 - Detección de valores extremos (outliers)
4. **Análisis por cuantiles**
 - Productos con ventas superiores al cuantil 75
 - Productos con ventas inferiores al cuantil 25
 - Comparación por categoría y producto

CONCLUSIONES POR SEGMENTO

Ventas totales por categoría



Aquí se muestra:

- Diferencias claras entre categorías
- Una o dos categorías concentran la mayor parte de las ventas

Se pudiera decir que:

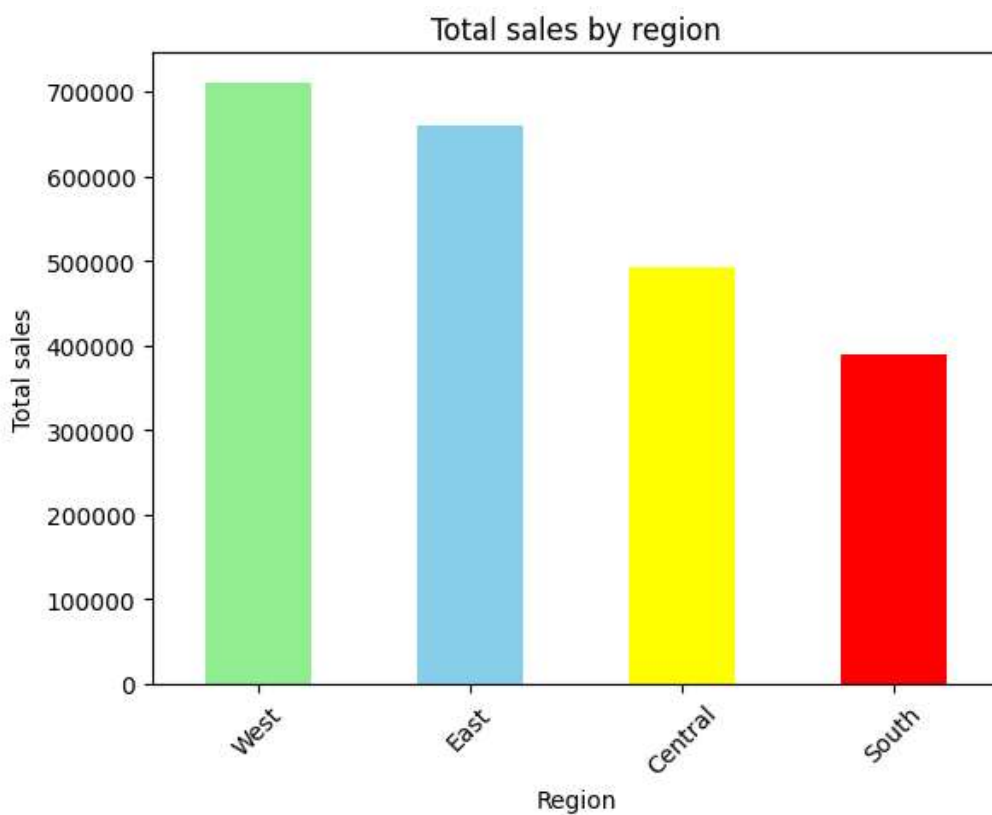
No todas las categorías aportan el mismo valor al negocio. Algunas categorías son claramente más relevantes en términos de ingresos, por ejemplo, technology la cual no solo lidera en ventas, sino que probablemente concentra los productos de mayor ticket promedio.

Conclusión

El negocio puede:

- Priorizar las categorías más fuertes
- Analizar si las categorías más débiles requieren ajustes en precio, promoción o catálogo

Ventas por Región



Aquí se muestra:

- Regiones con mayor volumen de ventas
- Regiones con menor participación

Se pudiera decir que:

El desempeño comercial varía según la región. Algunas zonas generan más ingresos que otras. Esto pudiera asomar la posibilidad que las zonas donde las ventas son menores (por

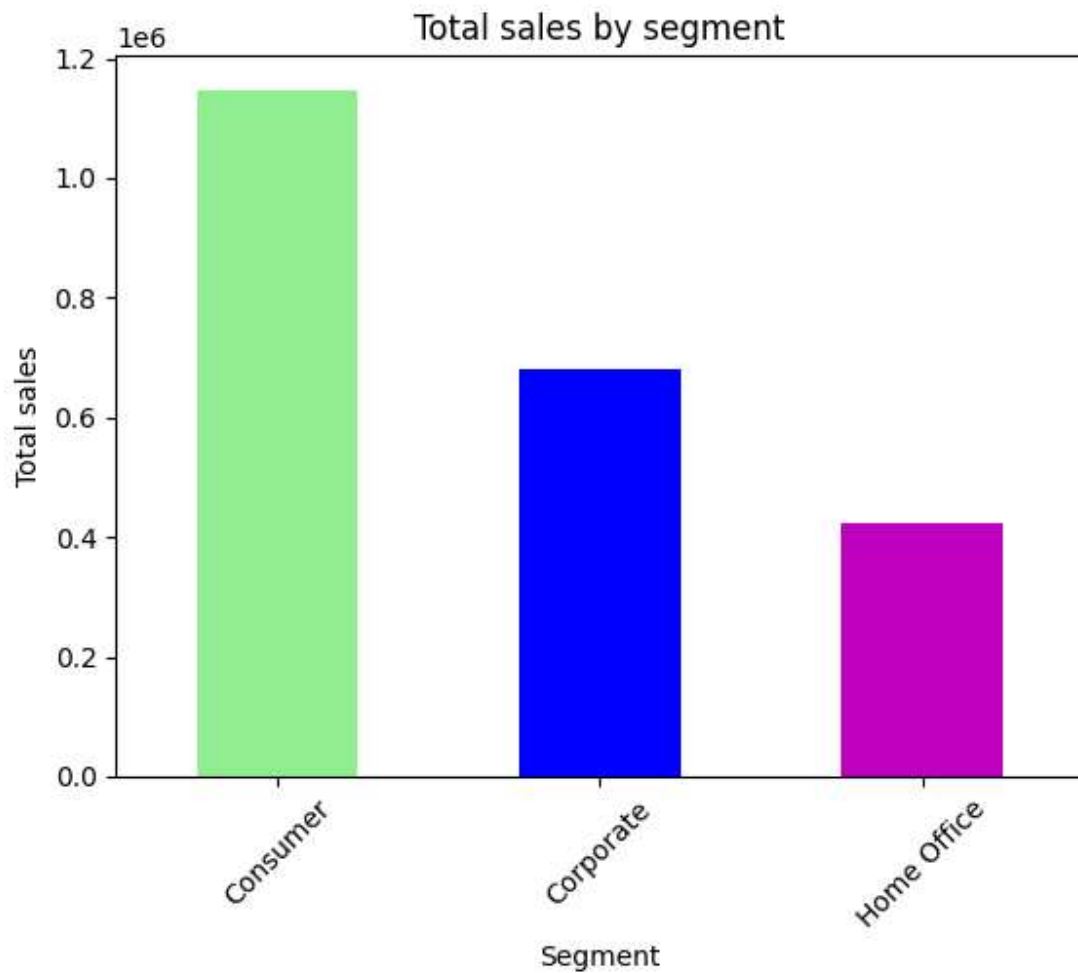
ejemplo, la zona central y south) pueden estar asociadas a factores externos como demanda, cobertura logística o estrategias comerciales locales que resten cuota de mercado.

Conclusión

Esto puede ayudar a:

- Enfocar esfuerzos comerciales en regiones clave
- Investigar por qué ciertas regiones venden menos (demanda, logística, cobertura)

Ventas por segmentación de clientes



Aquí se muestra:

- Diferencias claras entre tipos de clientes
- Un segmento destaca por generar más ventas

Se pudiera decir que:

No todos los clientes compran de la misma manera ni aportan el mismo valor. Algunos segmentos realizan compras más grandes o más frecuentes. Donde se muestra que los segmentos más bajos son Home Office (por debajo del 50%) seguido por Corporate, pero el de mayor impacto es Consumer

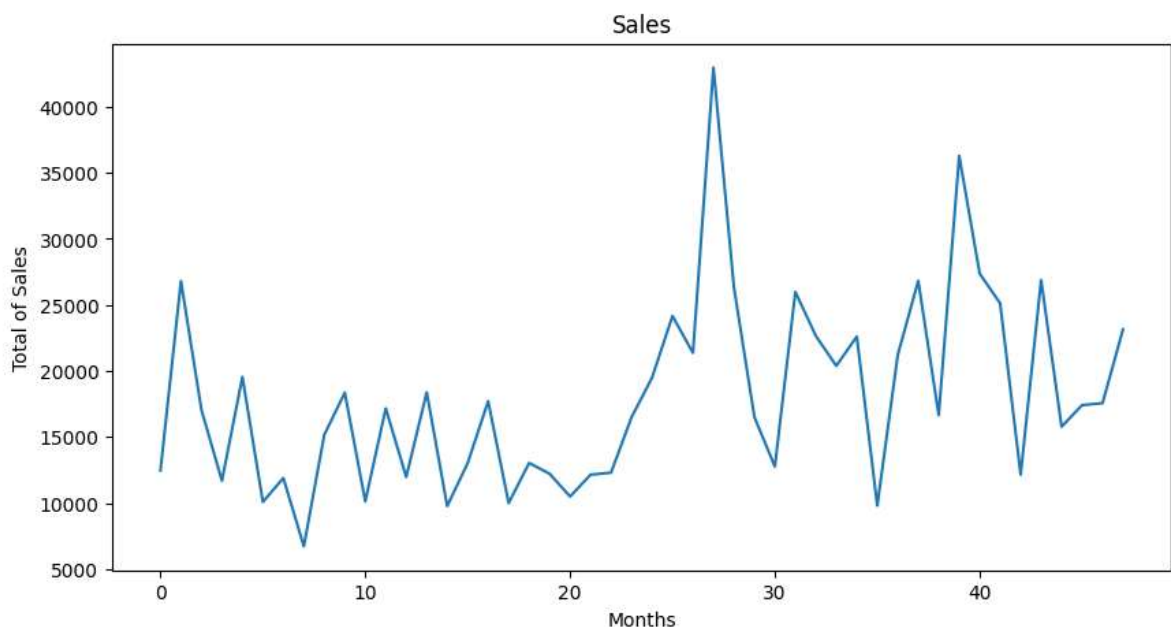
Pudiéramos decir que el segmento Consumer parece ser el motor del volumen de ventas, mientras que Corporate y Home Office podrían representar oportunidades de crecimiento con estrategias específicas.

Conclusión

El negocio puede:

- Identificar su segmento más rentable
- Diseñar estrategias específicas para cada tipo de cliente

Evolución de las ventas con el pasar del tiempo(mensual)



Aquí se muestra:

- Cambios en las ventas a lo largo del tiempo
- Meses con mayor y menor actividad

Se pudiera decir que:

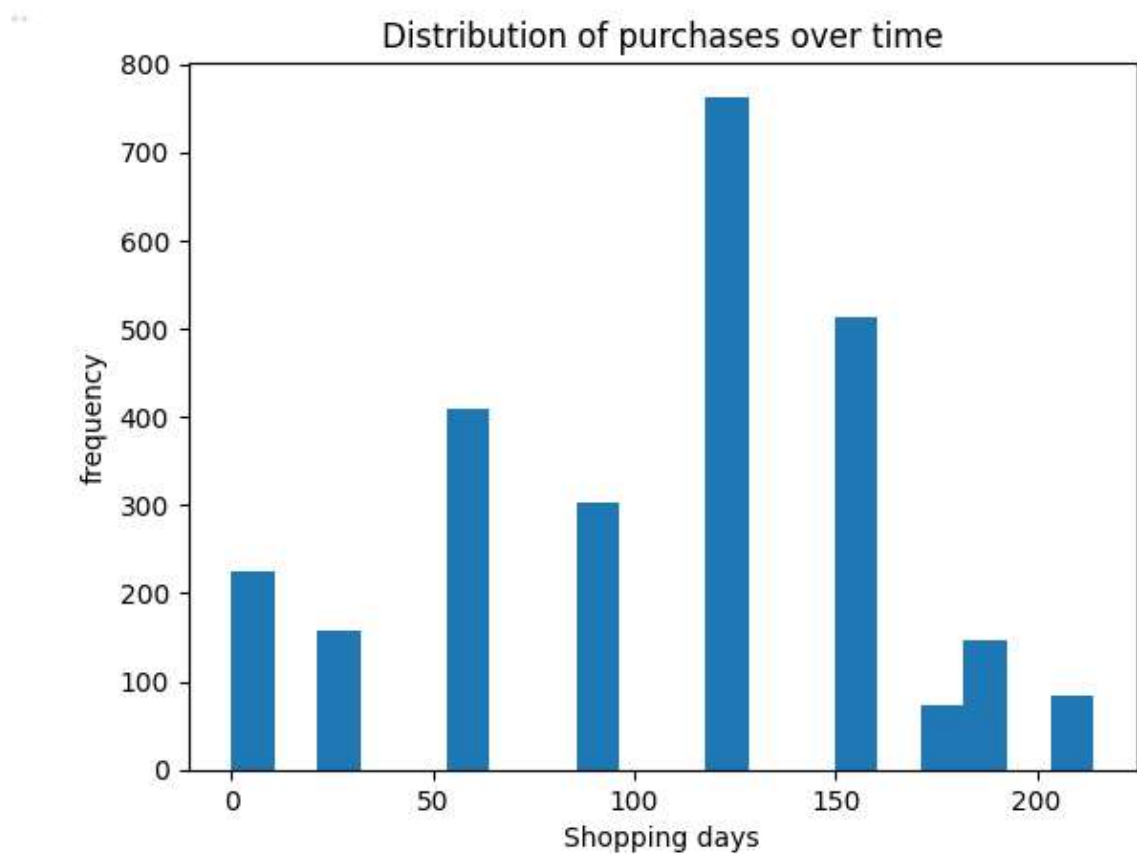
Las ventas no son constantes, existen períodos de mayor actividad y otros más bajos.

Conclusión

El negocio puede:

- Detectar posibles temporadas altas y bajas
- Planificar inventario y campañas con base en el comportamiento histórico

Tiempo de envió



Qué muestra el gráfico

- La mayoría de los pedidos se envían en pocos días
- Algunos pedidos tardan más tiempo en ser enviados

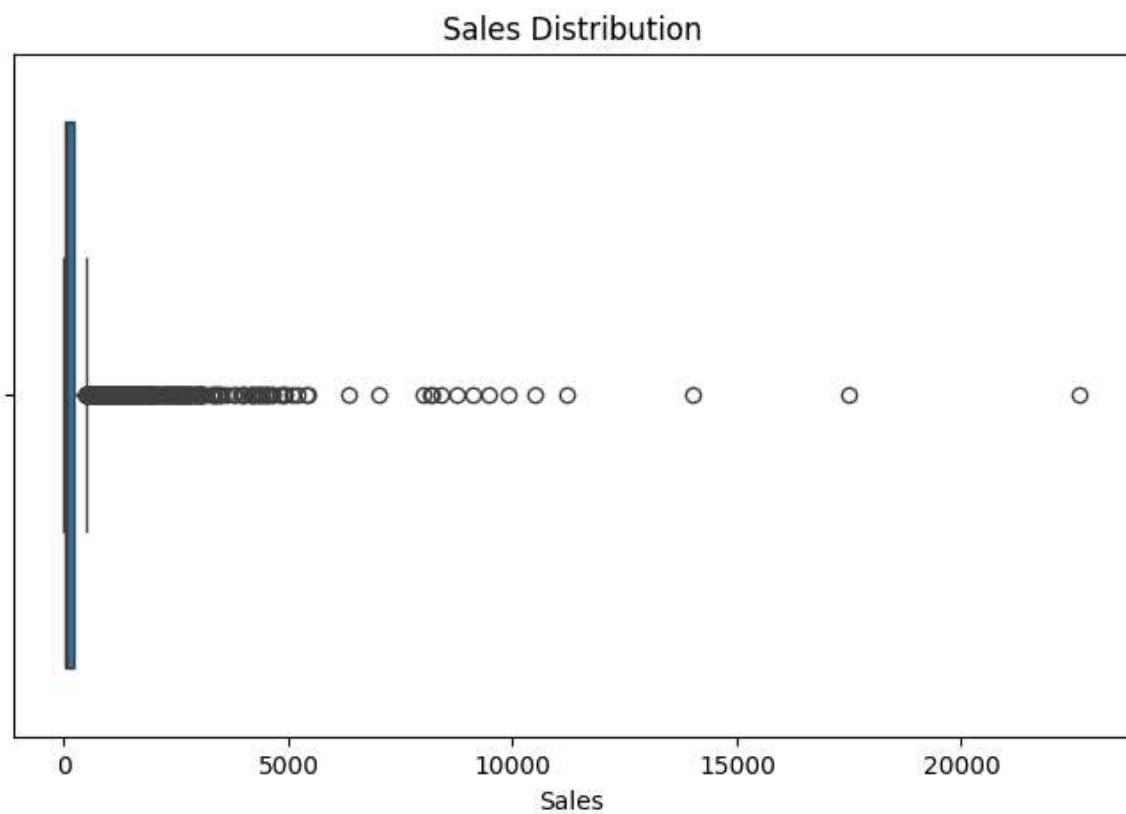
En general, el proceso de envío es razonablemente eficiente, aunque existen casos puntuales con mayor demora. Sería ideal evaluar si se pudiera aplicar los 5 por que

Conclusión

El negocio puede:

- Mantener el estándar actual de envío
- Analizar los casos con mayor demora para mejorar la experiencia del cliente

Distribución de ventas



Aquí se muestra

- Muchas ventas pequeñas concentradas cerca de cero
- Pocas ventas muy grandes (outliers)
- Una gran diferencia entre la mayoría de pedidos y unos pocos pedidos especiales

Se pudiera decir que:

La tienda realiza muchas ventas de bajo monto, pero una pequeña cantidad de pedidos representa ventas muy altas. Esto pudiera decir que el negocio atiende tanto a clientes minoristas como a clientes que hacen compras grandes. Se observa que Muchas ventas pequeñas sostienen el volumen, pocas ventas grandes impulsan el ingreso.

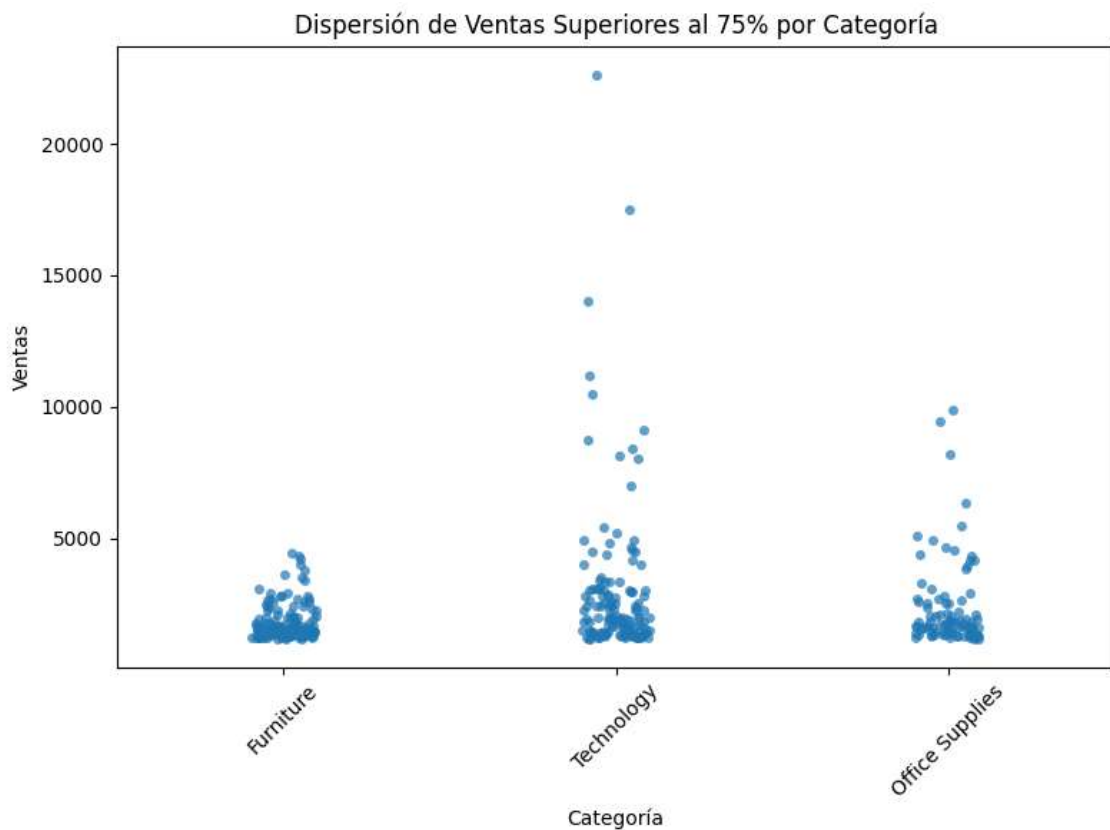
Conclusión

El negocio no depende de un solo tipo de cliente:

- Las ventas pequeñas sostienen el volumen
- Las ventas grandes impulsan el ingreso total

Otros gráficos

Productos con ventas por encima del cuantil 75%



Aquí se muestra

Un gráfico de dispersión donde se observan productos cuyos ingresos totales están por encima del cuantil 75, separados por categoría. Cada punto representa un producto con ventas acumuladas altas.

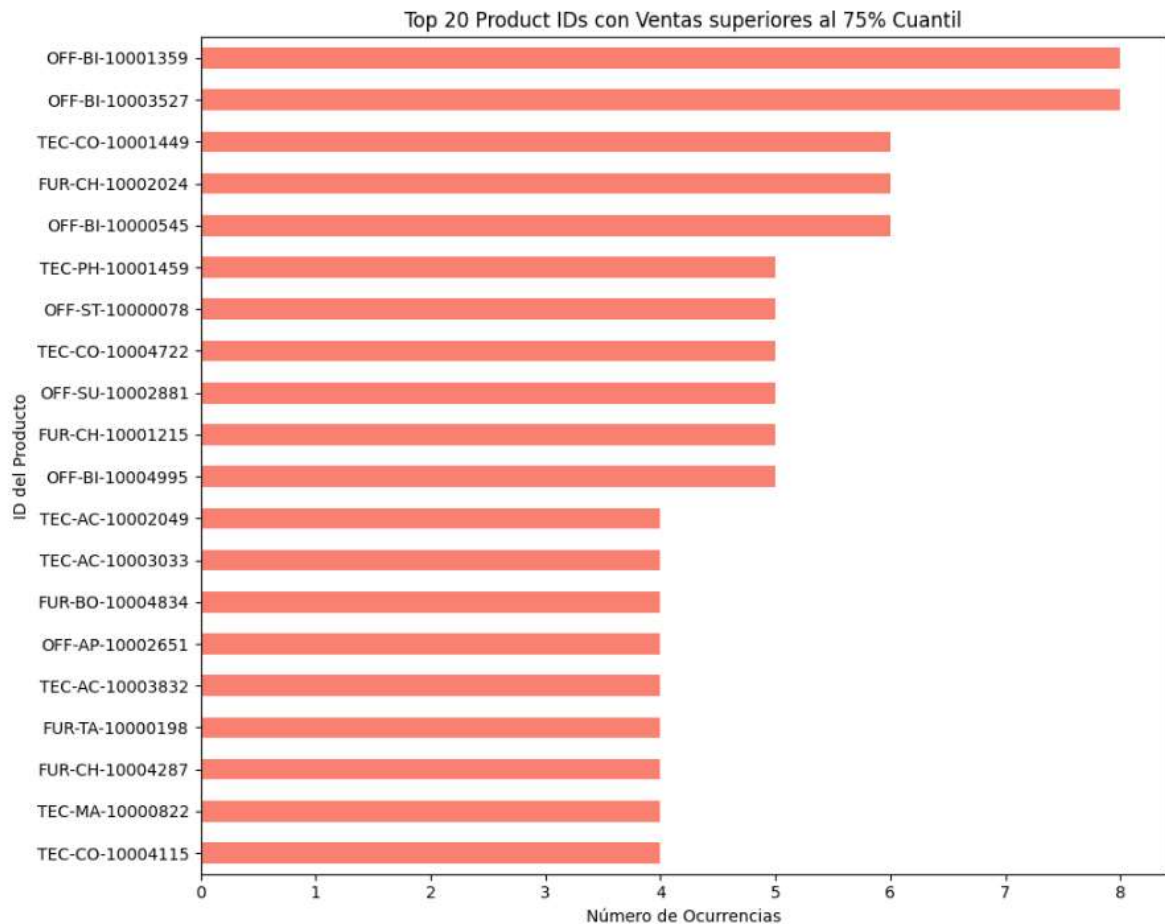
Se pudiera decir que:

El gráfico muestra que los productos de alto desempeño no están concentrados en una sola categoría, sino que aparecen distribuidos entre varias. Sin embargo, algunas categorías presentan mayor densidad de productos con ventas altas, lo que indica un mejor desempeño comercial general. Pero en el caso del segmento Technology tenemos el mayor número de outhliers, seguido por una gran brecha por Office Supplies

Conclusión

El negocio no depende de un único tipo de producto para generar altos ingresos. Existen categorías más fuertes, pero el éxito en ventas se reparte entre distintas líneas, lo que reduce el riesgo y abre oportunidades para potenciar varias áreas del catálogo.

Top 20 Product IDs con Ventas superiores al 75% Cuantil



Aquí se muestra

Un ranking de los 20 productos con mayores ventas acumuladas, es decir, aquellos que se encuentran claramente por encima del comportamiento promedio.

Se pudiera decir que:

Un número reducido de productos concentra una parte importante del ingreso total. Estos productos destacan de forma consistente frente al resto del catálogo. Además, estos productos como el OFF-BI-10001359 y OFF-BI-10003527 son los que generan más ingresos por ventas

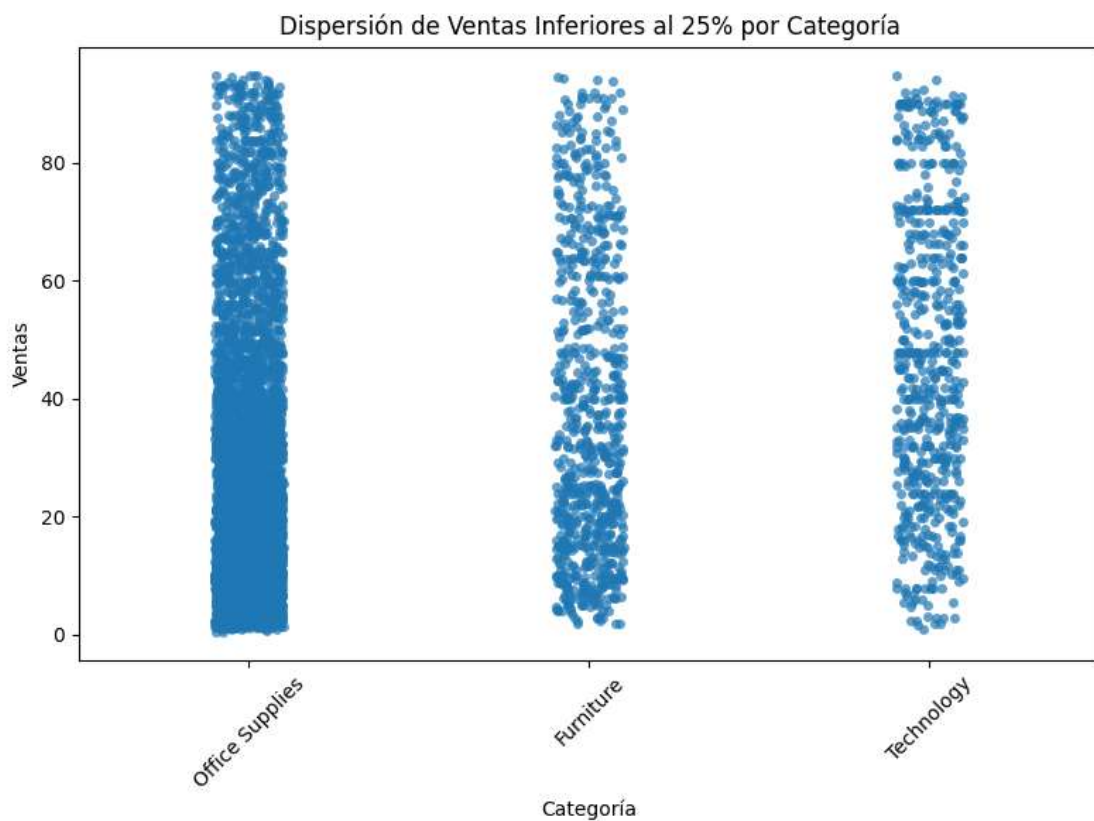
Conclusión

Estos productos representan activos clave para el negocio.

Pueden ser priorizados para:

- Estrategias de promoción
- Gestión de inventario
- Un análisis adicional de estos productos (precio, margen o demanda) permitiría entender por qué sobresalen.

Dispersión de Ventas Inferiores al 25% por Categoría



Aquí se muestra

Un gráfico de dispersión de productos cuyo total de ventas se encuentra por debajo del cuantil 25, segmentados por categoría.

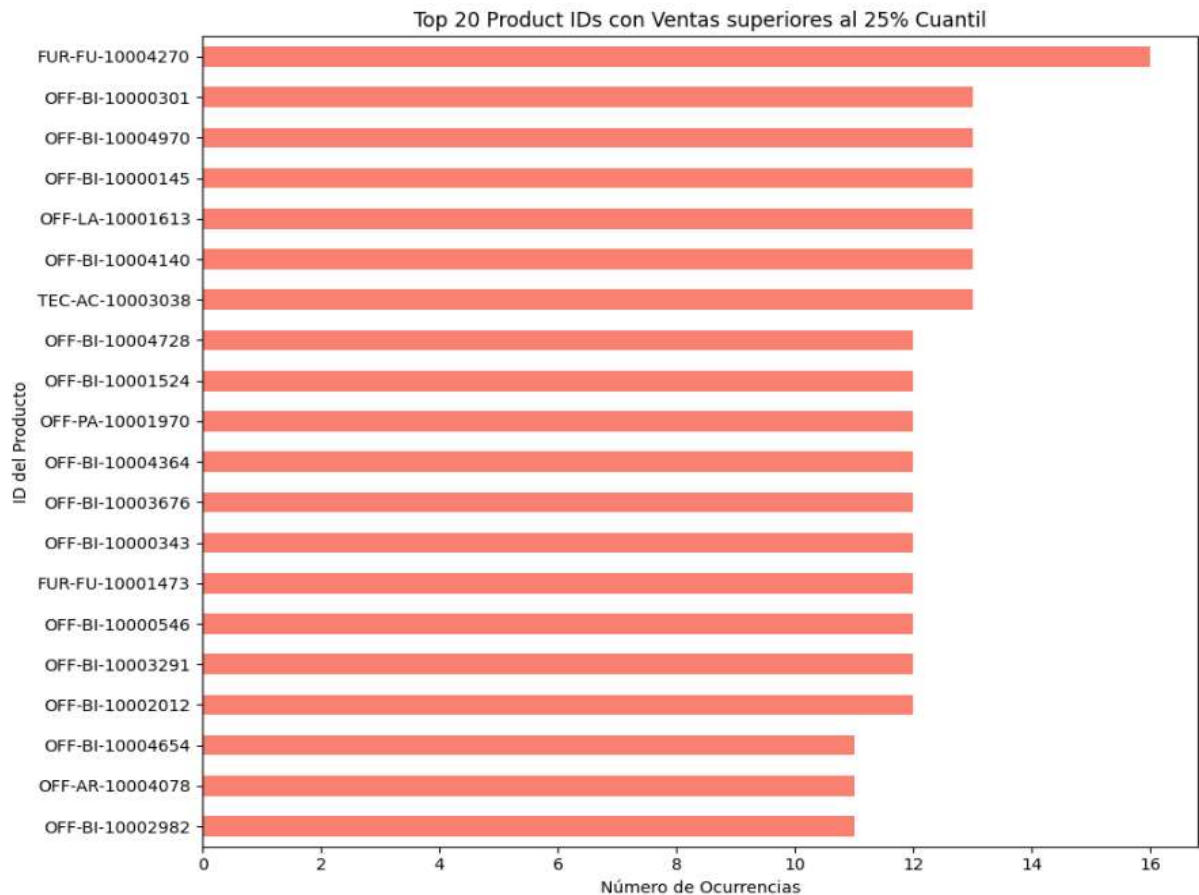
Se pudiera decir que:

Se observa que las ventas bajas también están distribuidas entre distintas categorías, sin concentrarse exclusivamente en una sola. Esto sugiere que el bajo desempeño no depende únicamente del tipo de producto, sino posiblemente de otros factores como demanda o visibilidad debido que el bajo desempeño no depende únicamente del tipo de producto

Conclusión

Existen productos con bajo rendimiento en varias categorías. Esto no implica que las categorías sean débiles, sino que ciertos productos individuales no están generando suficiente volumen de ventas y podrían requerir revisión.

Top 20 Product IDs con Ventas Inferiores al 25% Cuantil



Aquí se muestra

Un ranking de los 20 productos con menor volumen de ventas acumuladas dentro del dataset.

Se pudiera decir que:

Estos productos presentan un desempeño comercial claramente inferior al resto del catálogo. Su contribución al ingreso total es baja y no tienen un impacto significativo en las ventas globales.

Conclusión

Estos productos podrían ser candidatos para:

- Revisión de estrategia comercial
- Ajustes de precio
- Promociones específicas
- Evaluación de continuidad en el catálogo

LIMITACIONES Y PRÓXIMOS PASOS

1. Se pudiera calcular la relación de los cuantiles con las regiones a ver si existe una relación entre ellas y las bajas ventas
2. Se pudiera calcular la rentabilidad de las ventas, pero para esto se necesitaría el costo de reposición, logística, gastos administrativos, impuestos y otros factores, pero para esto se necesita definir tanto la ubicación y el tamaño de organización adicionalmente obtener el dataset del valor de reposición de cada producto y costo operacionales. Pero esto haría realizar una serie de gráficos que no es de importancia para la lógica de este ejercicio

No se calculó más en profundidad la evolución de las ventas con el pasar del tiempo debido que las columnas `order_data` y `ship_date` tienen muchos valores denominados NaT lo que puede deberse a pedidos incompletos o pedidos que aún no han sido enviados.

La eliminación de estos registros implicaría una reducción significativa del dataset, pasando de 9,789 filas a aproximadamente 2,700, es decir, una pérdida superior al 60% de la información disponible.

Eliminar una proporción tan alta de datos podría generar una interpretación sesgada de los resultados y afectar negativamente la representatividad del análisis.

Dado que en este análisis se basa en otros aspectos no se necesita tener valores de calidad en columnas de fechas, se decidió conservar estos registros, ya que su eliminación también impactaría otras variables clave para el análisis de ventas y comportamiento del cliente donde si se necesita una gran cantidad de data

3. Se pudiera filtrar aún más los outliers si se tuviera la columna de precio unitario o la columna cantidades de cada producto debido que solo se tiene la columna ventas que pudiera estar haciendo referencia al valor de la venta total

CONCLUSIONES GENERALES:

El modelo de ventas de la Superstore se sostiene sobre un alto volumen de ventas de bajo monto, complementado por un conjunto reducido de productos con ventas significativamente altas.

Este comportamiento sugiere:

- Estabilidad en el flujo de ingresos
- Dependencia estratégica de ciertos productos clave
- Oportunidades claras de optimización en productos de bajo desempeño

HERRAMIENTAS UTILIZADAS

- Lenguaje de programación: Python
- Librerías principales:
 - Pandas -para el análisis de la data
 - Matplotlib -para la visualización de los resultados
 - Seaborn -para la visualización de los resultados
 - Os -se usó para interactuar con el sistema operativo para manejar las rutas de los archivos
 - Textwrap -para colocar una nota grande al lado de un grafico
- Editor de código: colab
- IA:
 - ChapGPT: para generar la problemática y lógica de negocio
 - Gemini 2.5 incorporado en colab para correcciones de errores, mejora de código, soluciones de problemas, etc

ESTADO DEL PROYECTO:

Proyecto finalizado para análisis exploratorio y portafolio.