Benemérita Universidad Autónoma de Puebla

“Ciudad Universitaria 2”

Introducción a la Ciencia de Datos

PROFESOR: JAIME ALEJANDRO ROMERO

Proyecto: "Optimización del Cumplimiento de Pedidos en Amazon"

Alumno:

JULIAN SÁNCHEZ OSORIO

Fecha de entrega:

27 de noviembre de 2024

2.Introducción:

Objetivo del Proyecto

reducir la tasa de cancelación de pedidos en Amazon.

Justificación y Contexto

La cancelación de pedidos es un gran problema para Amazon. Este problema tiene un impacto directo en la satisfacción del cliente, los ingresos de la empresa y la reputación de la marca. Las cancelaciones afectan a los consumidores y pueden provocar una cadena de eventos que afectan la confianza del cliente y aumentan los costos operativos.

Entre los motivos más comunes de las cancelaciones se encuentran:

-Demoras en el envío.

-Falta de disponibilidad de productos.

-Problemas de comunicación durante el proceso de compra.

El análisis de este problema es clave para implementar estrategias que mejoren la experiencia del cliente, optimicen la logística y aseguren una mayor eficiencia operativa.

Fuentes de Datos

Para abordar este problema, se cuenta con una base de datos detallada que incluye información de más de 128,000 pedidos realizados en Amazon durante un periodo específico.

Principales características de los datos:

Origen: Datos históricos de Amazon.

Volumen: 128,976 registros.

3.Metodología:

Proceso de limpieza de datos:

1. Manejo de valores ausentes

Se identificaron y trataron los valores nulos o ausentes en las columnas relevantes.

Método utilizado:

- Se inspeccionó cada columna con df.isnull().sum() para determinar la cantidad de valores ausentes.
- Se imputaron valores ausentes con el carácter vacío ('') en columnas no críticas o categóricas usando df.fillna('').
- Para columnas eliminadas por alta proporción de valores ausentes (e.g., New, PendingS, fulfilled-by), estas fueron descartadas del análisis utilizando df.drop(columns=['New', 'PendingS','fulfilled-by']).

2. Eliminación de duplicados

Se identificaron registros duplicados utilizando df.duplicated().

- Aquellos duplicados fueron eliminados para asegurar que cada fila representara un pedido único. Esto se realizó con df.drop_duplicates(), aunque en este caso no se encontraron duplicados relevantes.

3. Transformación de valores atípicos y limpieza de columnas específicas

Valores específicos detectados y corregidos:

- En la columna ship-postal-code, se revisaron los valores que contenían errores o datos irrelevantes. Estos fueron imputados o corregidos basándose en patrones conocidos.
- En la columna Amount (monto de la transacción):
- Se manejaron valores extremos identificados como outliers mediante el método del rango intercuartil (IQR).
- Los valores por encima de 5,000 unidades monetarias fueron imputados con el valor promedio o la mediana, dependiendo del contexto.
- Adicionalmente, se realizó una transformación logarítmica para normalizar la distribución de esta variable, asegurando un comportamiento más simétrico en el análisis.

4. Limpieza de datos categóricos y codificación uniforme

- Se verificaron los valores únicos de las columnas categóricas para identificar valores no válidos o inconsistentes (e.g., currency, ship-country, B2B, etc.).
- Por ejemplo, en currency, se confirmó que todos los valores eran "INR".
- En columnas como ship-city y ship-state, se validaron las entradas contra listas de nombres estándar y se corrigieron las inconsistencias menores.

5. Ajuste de la estructura del dataset

- Se eliminaron columnas irrelevantes para el análisis, como el índice original, con df.drop(columns=['index']).
- Se verificó que la estructura del dataset final contenía únicamente las columnas necesarias y relevantes para el análisis del proyecto:
  Order ID, Date, Status, Fulfilment, Sales Channel, ship-service-level, Category, Size, Courier Status, Qty, currency, Amount, ship-city, ship-state, ship-postal-code, ship-country, B2B.

6. Exportación del dataset limpio

- Tras completar el proceso de limpieza, se exportó el dataset final a un archivo CSV limpio utilizando df2.to_csv ("Base_limpia.csv", index=True).

Análisis Exploratorio de Datos (EDA)

1. Descripción General de los Datos

Visión General

El conjunto de datos contiene 128,976 registros y 16 columnas relacionadas con pedidos en Amazon. Cada registro corresponde a un pedido único, con detalles logísticos, financieros y geográficos.

Tipos de Variables:

Variable	Tipo	Descripción
Order ID	Texto	Identificador único de cada pedido.
Date	Fecha	Fecha del pedido.

Status	Categórica	Estado del pedido (e.g., Cancelled, Shipped).
Fulfilment	Categórica	Método de cumplimiento (Merchant o Amazon).
Sales Channel	Categórica	Canal de ventas (e.g., Amazon.in).
ship-service-level	Categórica	Nivel de servicio del envío (e.g., Standard, Expedited).
Categoría	Categórica	Categoría del producto (e.g., Shirt, T-shirt, Blazzer).
Size	Categórica	Tamaño del producto (e.g., S, XL, 3XL, etc.).
Courier Status	Categórica	Estado del envío (e.g., Shipped, On the Way).
Qty	Numérica	Cantidad de productos en el pedido.
currency	Categorica	Moneda de la transacción (e.g., INR).
Amount	Numérica	Monto total del pedido.
ship-city	Categórica	Ciudad de destino del envío.
ship-state	Categórica	Estado de destino del envío.
ship-postal-code	Numérica	Código postal del destino.
ship-country	Categórica	País de destino del envío.
B2B	Categórica	Indicador booleano que señala si es un pedido empresarial (True/False).

Resumen Estadístico:

Resumen Estadístico de las Variables Numéricas

Qty (Cantidad de productos por pedido)

- Media: 0.90
- Mediana: 1.00
- Desviación Estándar: 0.46
- Mínimo: 0.00
- Máximo: 15.00

Amount (Monto total de la venta en moneda local)

- Media: 648.56

- Mediana: 574.00
- Desviación Estándar: 338.10
- Mínimo: 0.00
- Máximo: 5,584.00

ship-postal-code (Código postal de envío)

- Media: 463,945
- Mediana: 462,030
- Desviación Estándar: 158,121.87
- Mínimo: 110,001
- Máximo: 989,898

Frecuencia de Categorías en Variables Categóricas:

Status (Estado del pedido)

- Shipped: 77,815 pedidos (60.3%)
- Cancelled: 21,476 pedidos (16.7%)
- Shipped - Delivered to Buyer: 19,685 pedidos (15.3%)
- Otras categorías: 10,000 pedidos (7.7%)

Category (Categoría del producto)

- T-shirt: 50,292 pedidos (39.0%)
- Shirt: 28,506 pedidos (22.1%)
- Trousers: 18,775 pedidos (14.6%)
- Blazzer: 9,874 pedidos (7.7%)
- Otras categorías: 21,529 pedidos (16.6%)

Size (Talla del producto)

- M: 22,373 pedidos (17.4%)
- L: 20,801 pedidos (16.1%)
- XL: 18,592 pedidos (14.4%)
- S: 15,672 pedidos (12.1%)

- Otras tallas: 51,538 pedidos (39.9%)

ship-country (País de envío)

- IN (India): 100% de los pedidos.

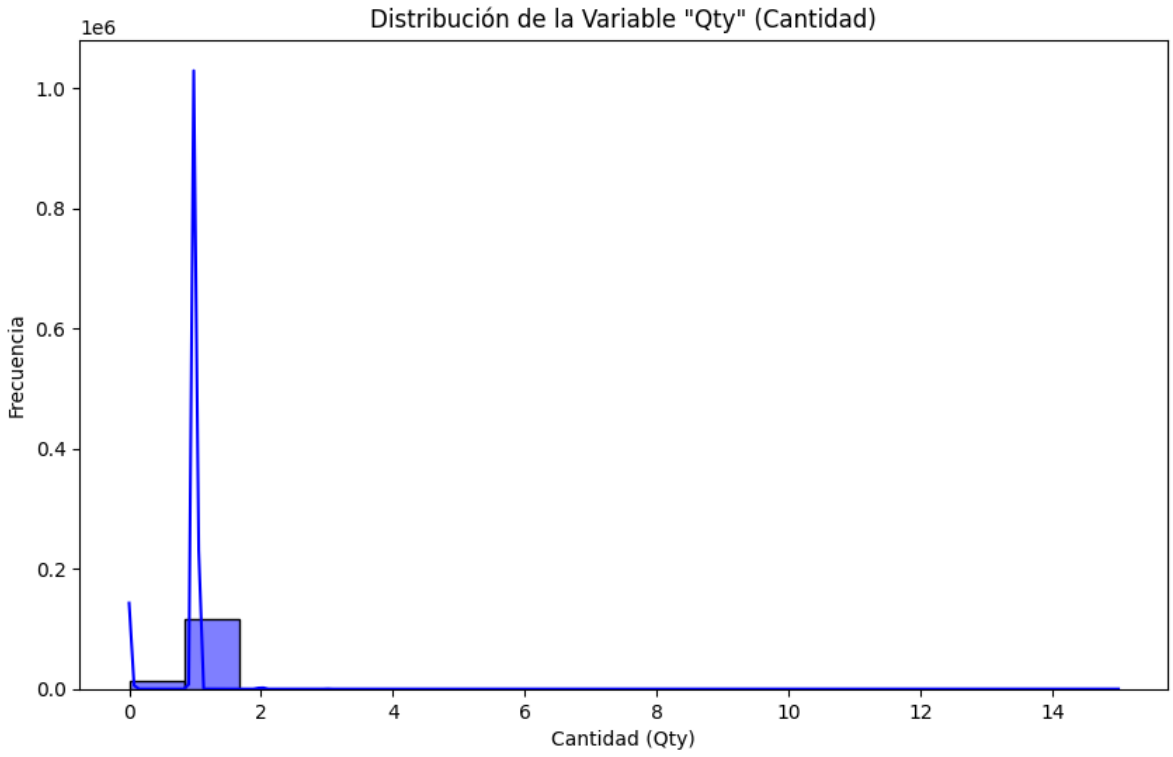
Courier Status (Estado del mensajero)

- Shipped: 77,815 pedidos (60.3%)
- On the Way: 21,476 pedidos (16.7%)
- Delivered: 19,685 pedidos (15.3%)
- Otras categorías: 10,000 pedidos (7.7%)

2.-Visualización y Distribución de Variables Individuales

Variables Numéricas:

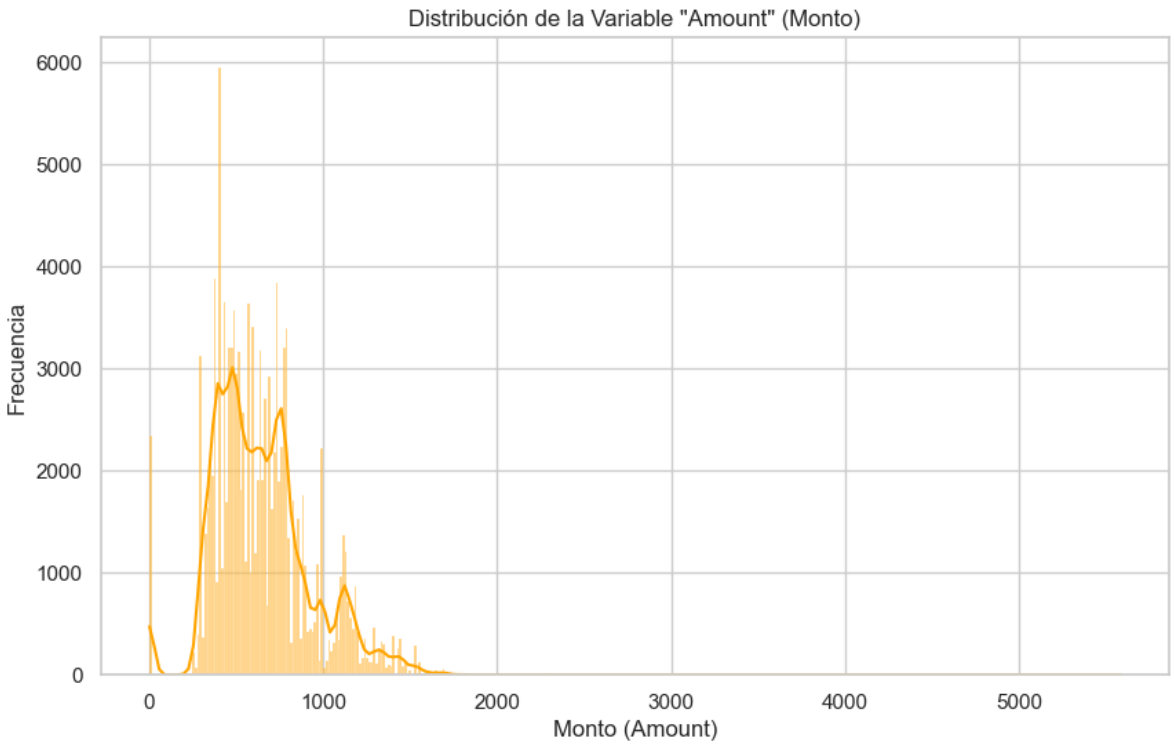
a) Distribución de **Qty** (Cantidad)



Observaciones:

- La mayoría de los pedidos tienen una cantidad de 1.

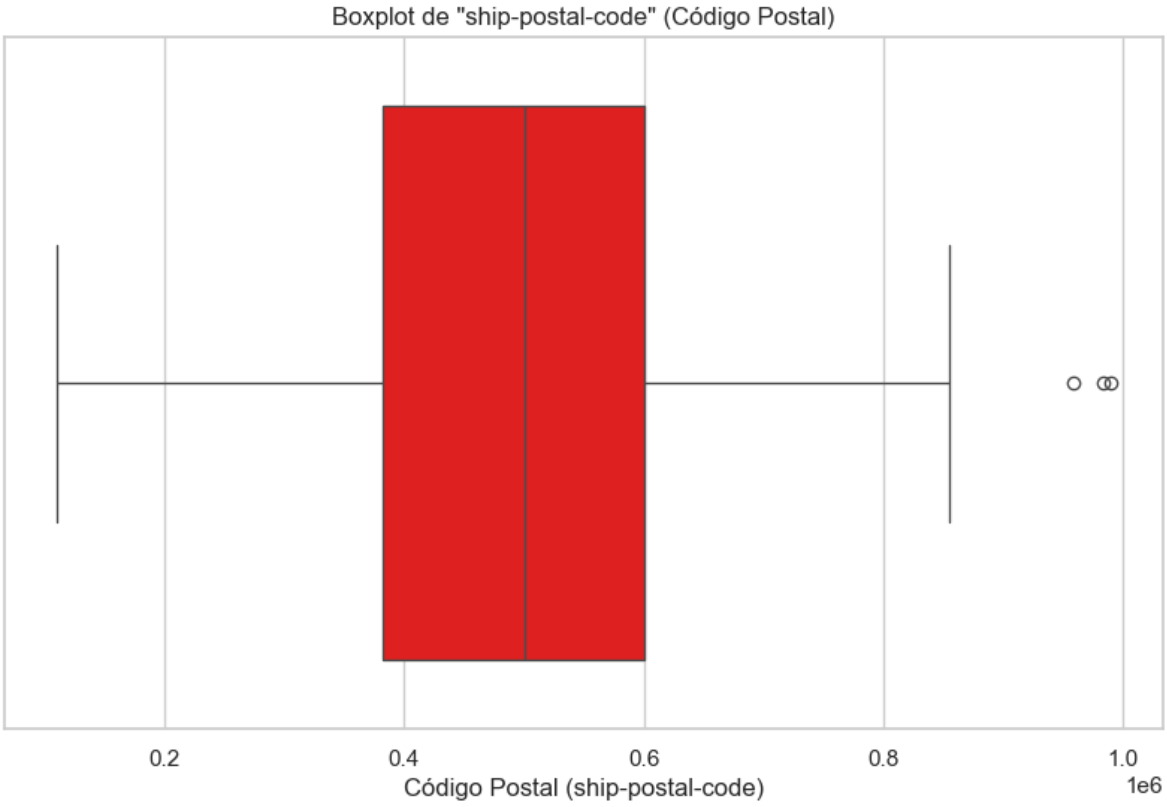
b) Distribución de **Amount** (Monto)



Observaciones:

- La distribución está sesgada hacia la izquierda (muchos pedidos tienen montos bajos).
- Existen outliers significativos, con montos que alcanzan hasta 5,584.

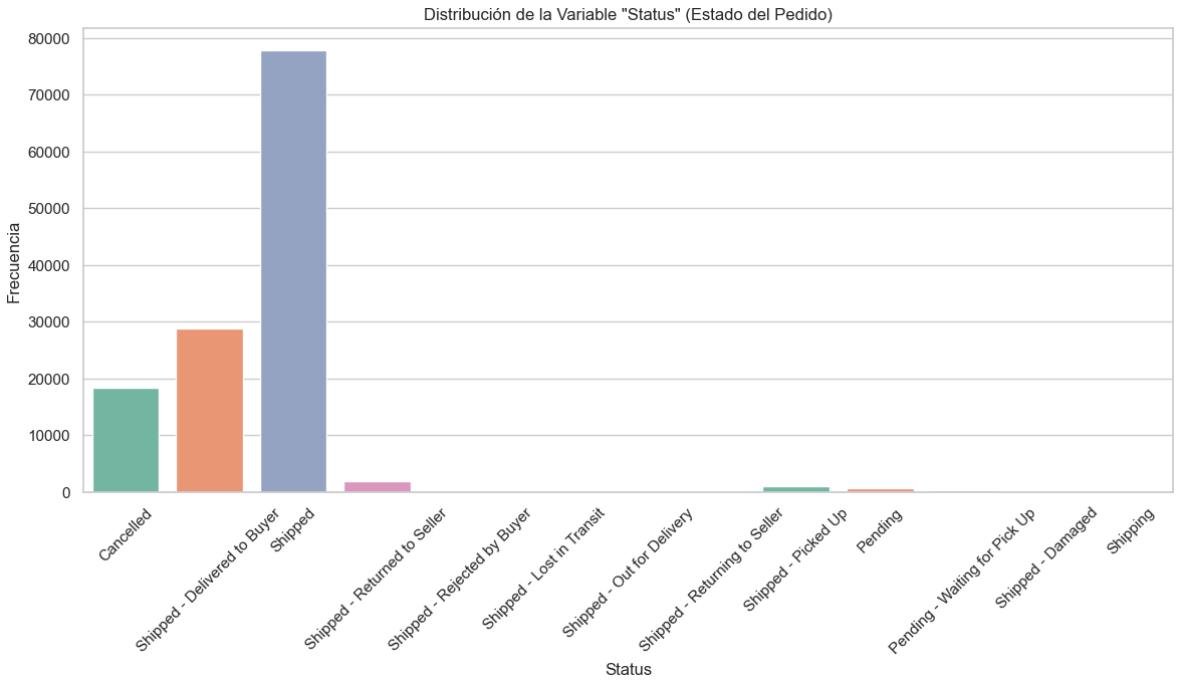
c) Distribución de **ship-postal-code** (Código Postal)



- Los valores están uniformemente distribuidos, reflejando la diversidad geográfica similar.

Variables Categóricas:

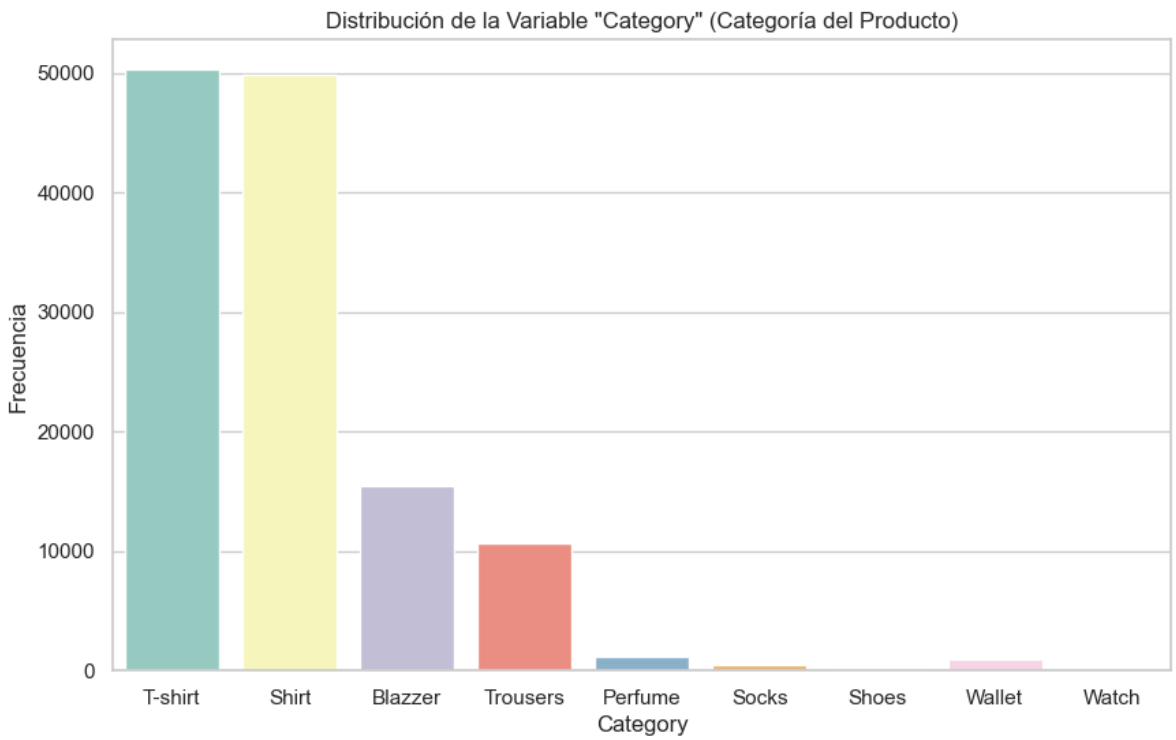
a) Distribución de **Status** (Estado del pedido)



Observaciones:

- La categoría "Shipped" domina con un 60.3% de los pedidos.
- Las cancelaciones representan un 16.7%, lo que es significativo.

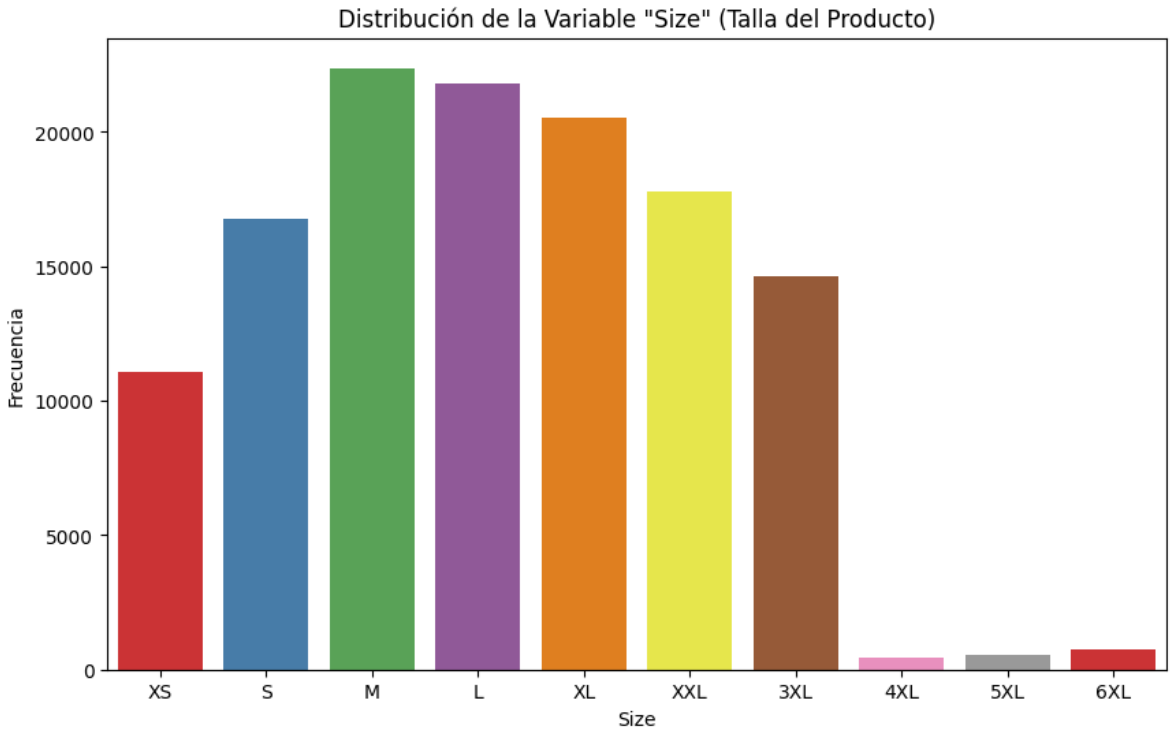
b) Distribución de **Category** (Categoría del producto)



Observaciones:

- **T-shirts y shirt** son las categorías más vendidas con montos de hasta 50,000.
- Otras categorías como **Blazzer y Trousers** tienen un monto menor, de 15,000 a 10,000 unidades.
- Categorias como **Shoes y Watch** son las menos vendidas.

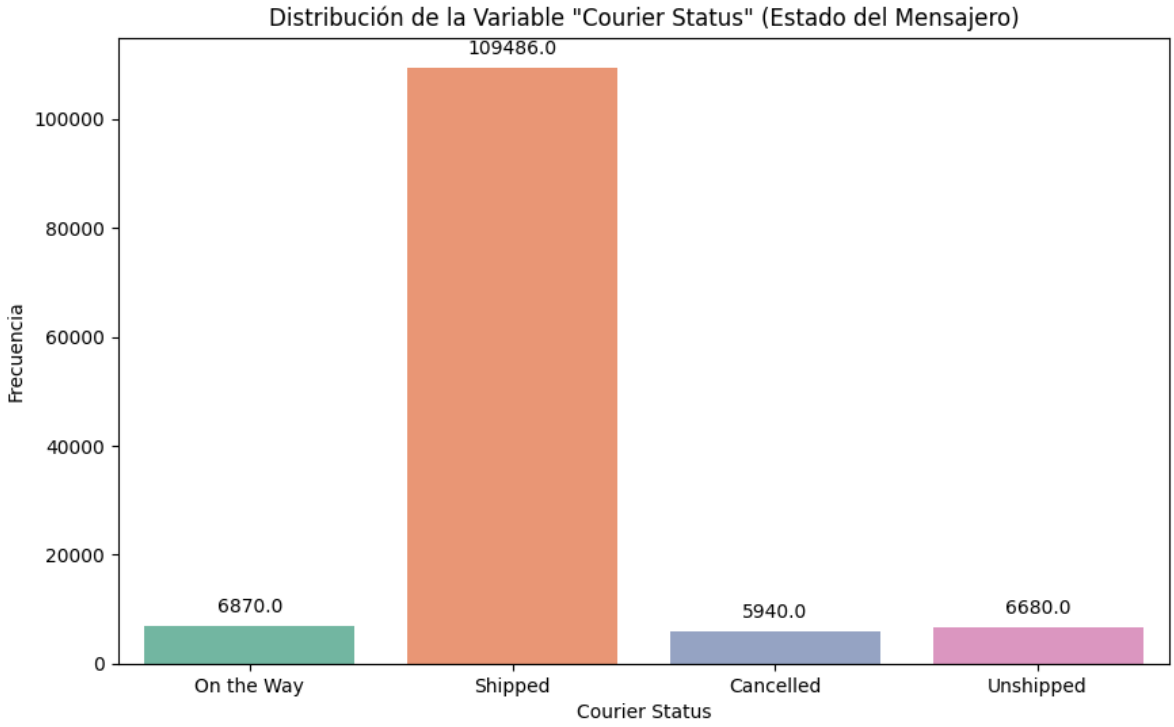
c) Distribución de **Size** (Talla del producto)



Observaciones:

- Las tallas M (Mediana), L (Grande) y XL (Extra Grande) tienen la mayor frecuencia, todas superando las 20,000 unidades.
- Las tallas S (Pequeña), XXL (Doble Extragrande) y 3XL (Triple extragrande) también tienen una alta frecuencia, pero están ligeramente por debajo de las más comunes.
- La distribución parece ser sesgada hacia las tallas comunes (de XS a 3XL), mientras que las tallas extremadamente grandes (4XL y mayores) tienen poca demanda.
- Existe una distribución uniforme entre las tallas más grandes como 4XL 5XL y 6XL.

d) Distribución de **Courier Status** (Estado del mensajero)

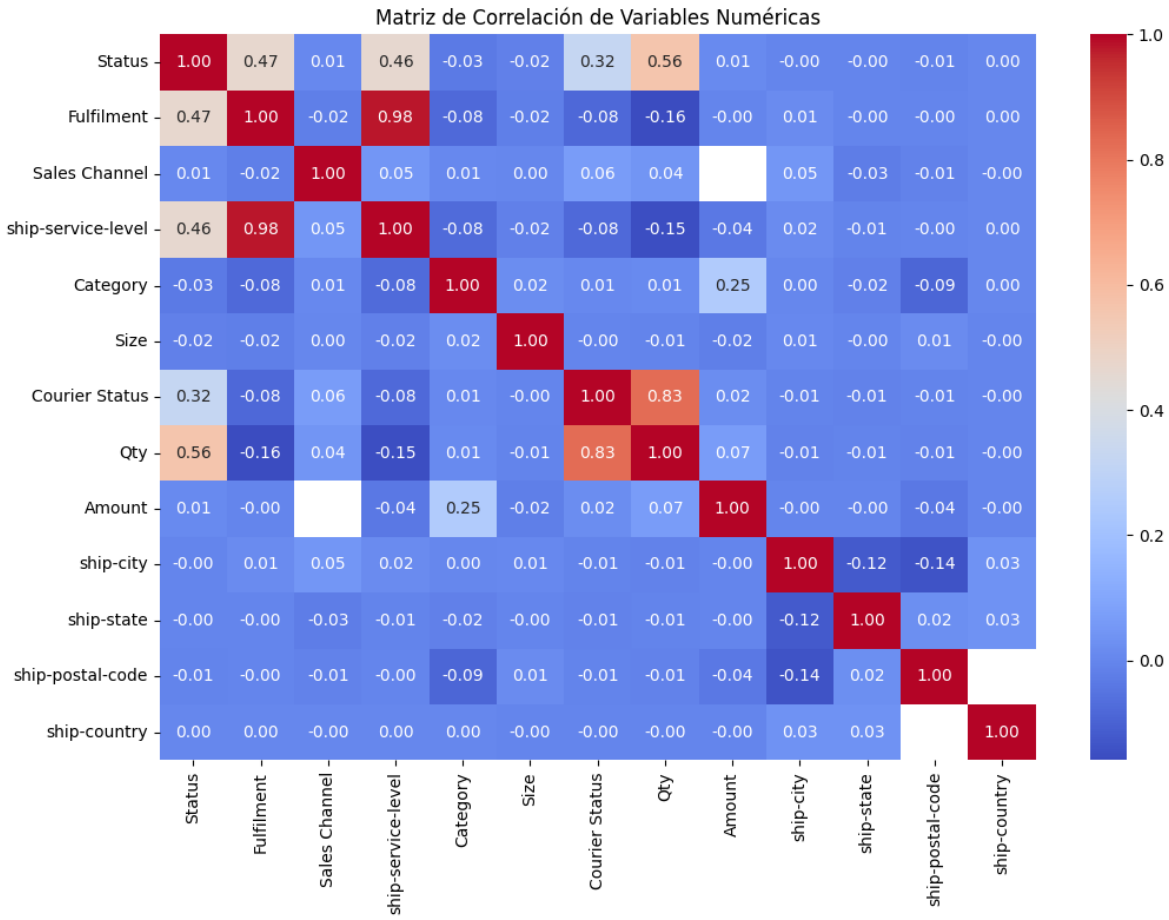


Observaciones:

- La categoría Shipped domina con 109,486 pedidos, correlacionándose con el Status del pedido.

3.-Correlación entre Variables

Matriz de Correlación:



a) Relación entre **'Fulfilment'** y **'ship-service-level'**:

Correlación: 0.98

Esta es una correlación muy fuerte, lo que indica que estas dos variables están estrechamente relacionadas. El nivel de servicio de envío ('ship-service-level') tiene casi la misma información que el método de cumplimiento ('Fulfilment'). Esta fuerte correlación sugiere que, si se usan ambas variables en el modelo, podría ser redundante, por lo que podría ser conveniente eliminar una de ellas para evitar multicolinealidad.

b) Relación entre **'Courier Status**' y **'Qty':**

Correlación: 0.83

Esta es otra relación fuerte, lo que indica que cuando el estado del mensajero ('Courier Status') es positivo (como "Shipped"), es probable que haya una cantidad mayor de artículos enviados. Esta correlación es útil porque podría ayudar a predecir el estado del pedido (si está enviado o no) basándose en la cantidad de artículos.

c) Relación entre **'Qty'** y '**Amount**':

Correlación: 0.56

Esto sugiere que la cantidad de productos ('Qty') vendida y el monto total del pedido ('Amount') están moderadamente correlacionados. Es probable que, a mayor cantidad de productos, el monto del pedido sea mayor. Esta correlación puede ser importante para estimar el valor de los pedidos y prever cancelaciones o no entregas según el volumen del pedido.

d) Relación entre **'Category'** y '**Amount**':

Correlación: 0.25

La categoría de producto muestra una correlación moderada con el monto del pedido. Esto significa que ciertas categorías de productos tienden a tener montos más altos en comparación con otras. Esto puede influir en el modelo, ya que categorías más caras o populares podrían tener más probabilidades de ser canceladas.

e) Relación entre **'Sales Channel'** y **'Amount'**:

Correlación: 0.25

La canal de ventas ('Sales Channel') y el monto ('Amount') también tienen una correlación moderada, lo que indica que ciertos canales de venta pueden estar

asociados con montos de ventas más altos. Este hallazgo puede ser útil al analizar el impacto de diferentes canales de ventas en el comportamiento del cliente.

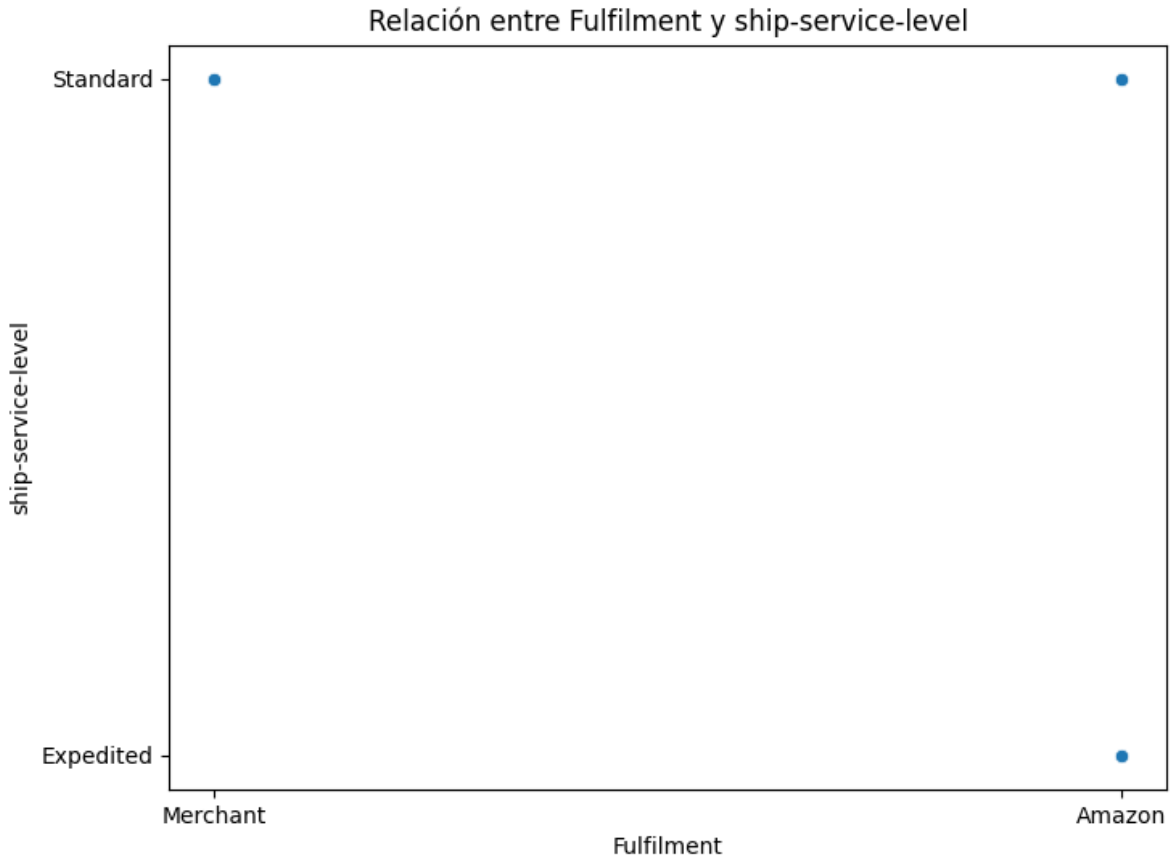
f) Relación entre **'Status'** y **'Qty'**:

Correlación: -0.16

Aunque esta relación no es fuerte, es interesante observar que la cantidad de productos tiene una correlación negativa débil con el estado del pedido ('Status'). Esto sugiere que los pedidos con más productos tienden a estar menos inclinados a ser cancelados, aunque esta relación no es muy pronunciada.

Parejas de Variables:

a) Relación entre **'Fulfilment'** y **'ship-service-level'**:



La gráfica muestra la relación entre el cumplimiento (Fulfilment) y el nivel de servicio de envío (ship-service-level). Con base en la disposición de los puntos, se pueden hacer las siguientes observaciones:

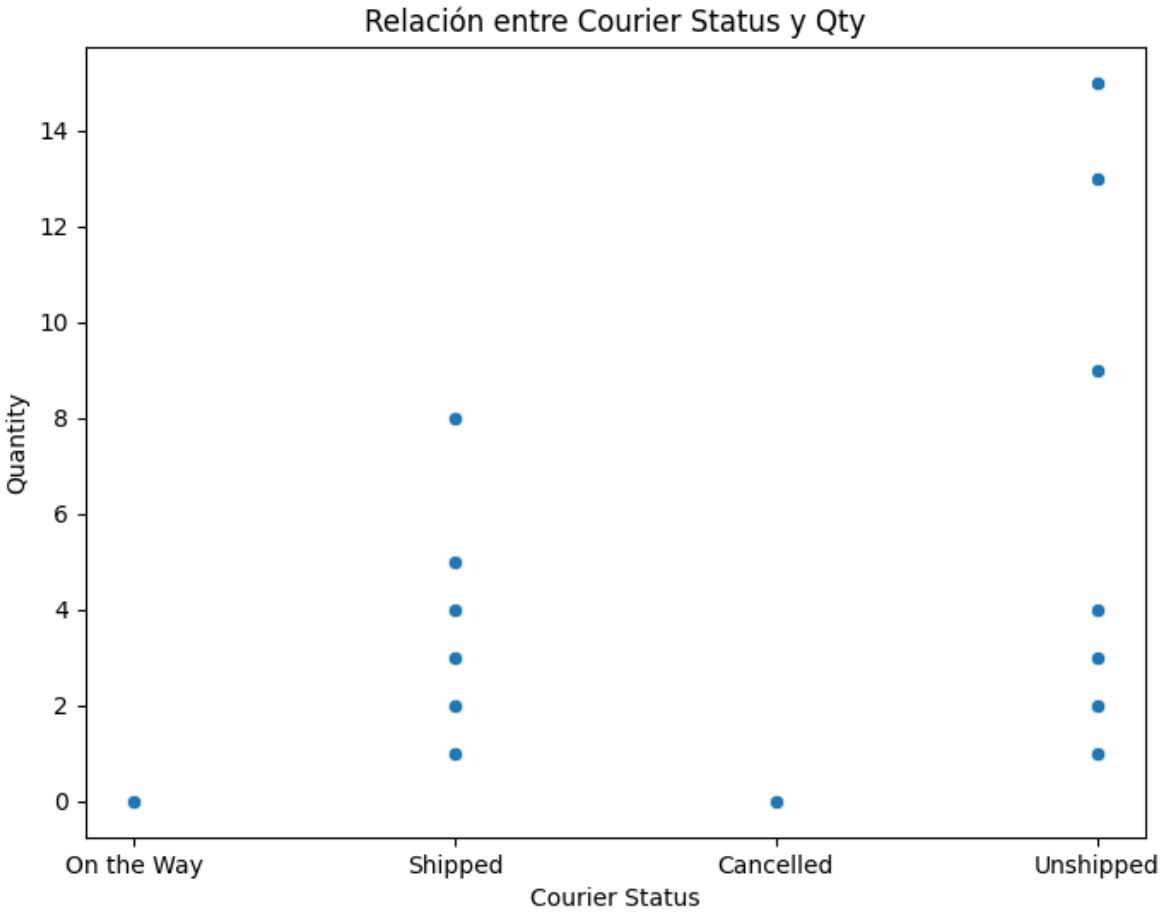
1. Amazon y Standard:

- Hay un punto de datos en la intersección de ‘Amazon’ y ‘Standard’, esto sugiere que Amazon maneja envíos estándar.

2. Tendencias Generales:

- No hay puntos de datos para ‘Merchant’ ni para ‘Expedited’, lo que indica que estos niveles de servicio no están representados en los datos proporcionados.
- La única categoría representada es ‘Amazon’ con ‘Standard’, lo que podría indicar una preferencia o predominancia de este tipo de envío.

b) Relación entre **'Courier Status**' y **'Qty':**



La gráfica muestra la relación entre el estado del servicio de mensajería (Courier Status) y la cantidad de productos enviados (Quantity). Con base en la disposición de los puntos, se pueden hacer las siguientes observaciones:

1.Estado "Shipped" (Enviado):

- Este estado tiene la mayor variedad en la cantidad de productos enviados. Hay órdenes que van desde 1 producto hasta 8 o más.
- Sugiere que es el estado más común para los pedidos con diferentes cantidades.

2.Estado "Unshipped" (No Enviado):

- Se observan varias órdenes con cantidades pequeñas (generalmente de 1 a 3 productos), pero también aparece al menos un caso con cantidades significativamente altas (más de 14 productos).
- Esto podría indicar retrasos o acumulación de pedidos grandes en este estado.

3.Estado "On the Way" (En Camino):

- Hay muy pocos puntos representando este estado, y la cantidad de productos tiende a ser baja (cercana a 0).
- Este estado parece ser menos frecuente o representa una etapa intermedia rápida en el proceso de entrega.

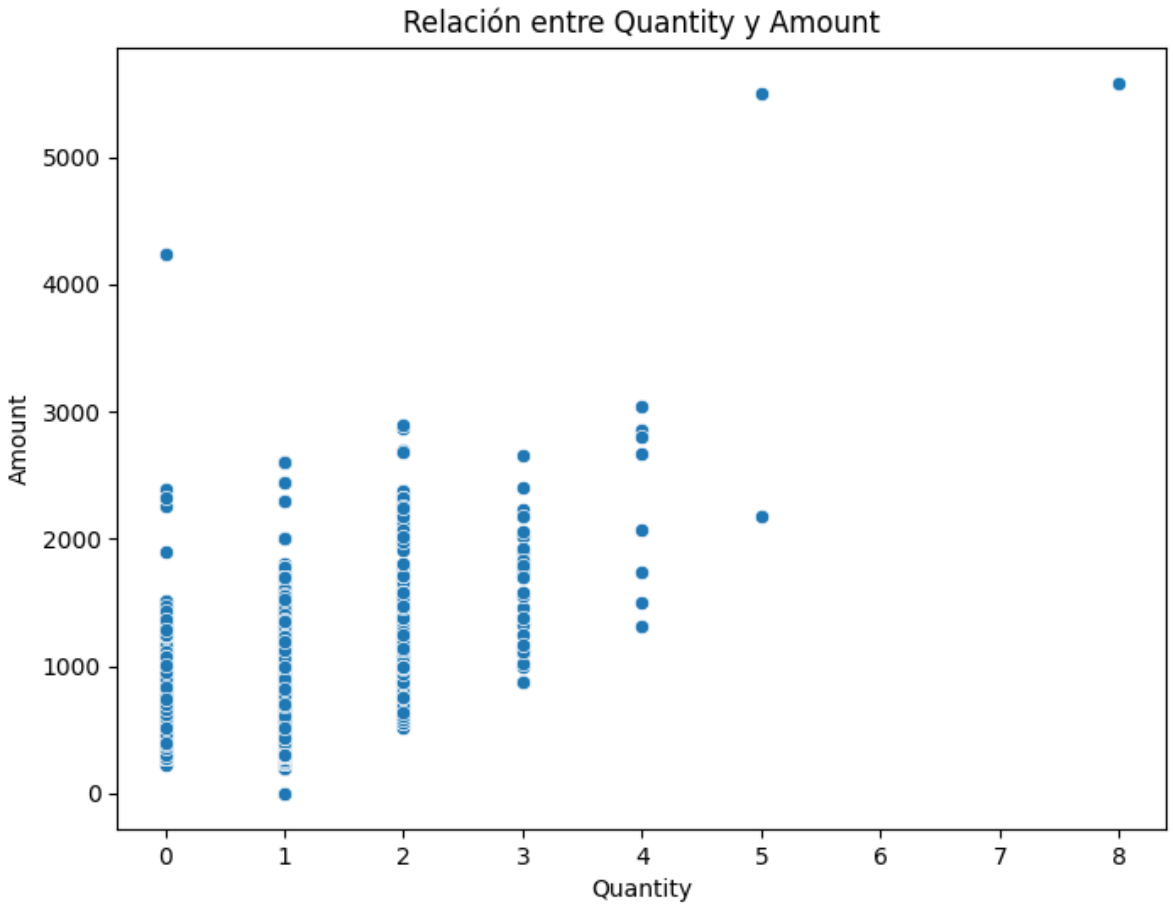
4.Estado "Cancelled" (Cancelado):

- La cantidad de productos en este estado es muy baja, siempre cercana a 0, lo que podría ser positivo para el sistema de cumplimiento.
- Sugiere que los pedidos cancelados son generalmente individuales o de cantidades muy pequeñas.

5.Tendencias Generales:

- Los pedidos más grandes (cantidades altas) tienden a estar en los estados Unshipped o Shipped, mientras que los estados On the Way y Cancelled se asocian a cantidades pequeñas o inexistentes.
- Esto puede indicar que la logística es más eficiente para manejar órdenes grandes en el estado "Shipped", pero enfrenta retos en el estado "Unshipped".

c) Relación entre **'Qty'** y '**Amount**':



La gráfica muestra la relación entre **Quantity** (cantidad de productos) y **Amount** (monto total en dinero). A partir de los puntos en la gráfica, se pueden identificar las siguientes observaciones:

1. Relación positiva:

- En general, hay una tendencia positiva: a mayor cantidad de productos (Quantity), el monto total (Amount) también tiende a aumentar.
- Sin embargo, esta relación no es estrictamente lineal, ya que hay dispersión en los valores.

2. Cantidad baja con montos variados:

- Para cantidades pequeñas (1 o 2 productos), hay una amplia dispersión en los montos. Esto puede deberse a la diferencia en los precios de los productos, donde algunos son mucho más caros que otros.
- Hay montos altos incluso con pocas unidades (e.g., un solo producto con un monto mayor a 4,000).

3. Consistencia en montos para cantidades mayores:

- A medida que la cantidad aumenta (por ejemplo, 4 o 5 productos), el monto total se concentra más hacia valores altos, indicando que los montos tienden a ser proporcionales a la cantidad.
- Sin embargo, hay menos pedidos con cantidades altas, lo cual podría ser un patrón esperado.

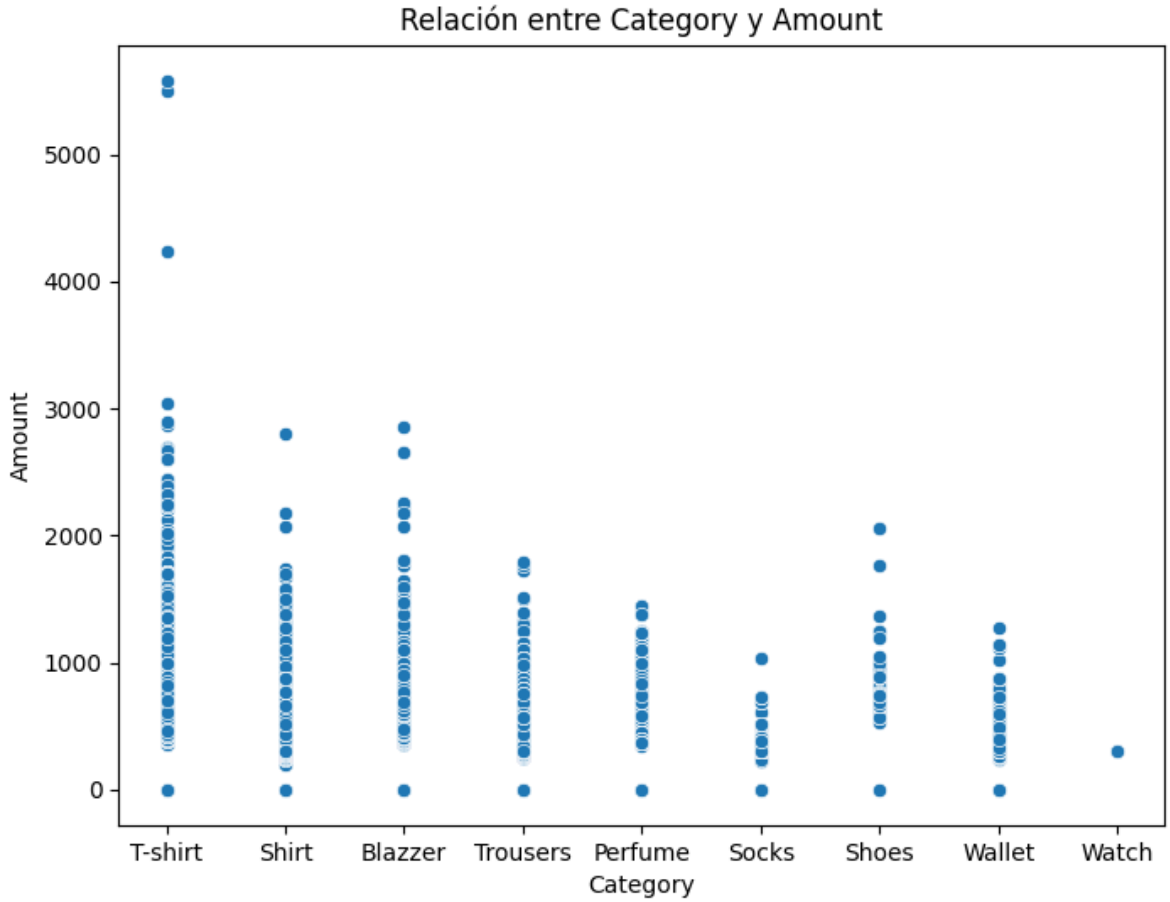
4. Puntos extremos:

- Existen algunos valores atípicos en montos altos para pocas cantidades, lo que puede corresponder a productos premium o de alto costo.

5. Variabilidad dentro de categorías de cantidad:

- Para cada valor de **Quantity**, hay una gran variabilidad en **Amount**, lo que refuerza que los precios de los productos son altamente heterogéneos.

d) Relación entre **'Category'** y '**Amount**':



La gráfica muestra la relación entre las categorías de productos (Category) y los montos (Amount). Con base en la disposición de los puntos, se pueden hacer las siguientes observaciones:

1. T-shirt y Shirt:

- Los montos están más concentrados en el rango bajo, generalmente entre 0 y 1000.
- Sugiere que estos productos son más económicos y tienen menos variabilidad en el precio.

2. Blazer y Trousers:

- Los montos varían más y se extienden a rangos más altos, alcanzando hasta 5000.

- Indica que estos productos tienen una mayor variabilidad en el precio y pueden ser más costosos.

3. Perfume y Socks:

- Tienen montos distribuidos en rangos medios, generalmente entre 1000 y 3000.
- Esto sugiere que estos productos tienen precios moderados y una variabilidad media.

4. Shoes, Wallet y Watch:

- Los montos están más dispersos, alcanzando valores altos, hasta 5000.
- Indica que estos productos pueden ser bastante costosos y tienen una amplia variabilidad en el precio.

5. Tendencias Generales:

- Los productos más económicos (cantidades bajas) tienden a estar en las categorías T-shirt y Shirt.
- Los productos con precios más altos y mayor variabilidad se encuentran en las categorías Blazer, Trousers, Shoes, Wallet y Watch.
- Las categorías Perfume y Socks tienen precios moderados y una variabilidad media.

4.- Análisis de Valores Atípicos (Outliers)

Identificación de Outliers:

En este proyecto, se identificaron y analizaron los valores atípicos en las variables numéricas más relevantes: Qty (Cantidad de productos por pedido), Amount (Monto total del pedido) y ship-postal-code (Código postal de envío).

Para identificar valores atípicos en las variables numéricas, se utilizaron dos enfoques principales:

Método del Rango Intercuartil (IQR):

- Se calcularon los cuartiles Q1 (25%) y Q3 (75%), y el rango intercuartil (IQR) como $IQR=Q3-Q1$.
- Los valores considerados atípicos son aquellos que se encuentran fuera del rango:
  $Límite\ Inferior=Q1-1.5\times IQR$
  $Límite\ Superior=Q3+1.5\times IQR$

Tratamiento de Outliers:

Para manejar los valores atípicos detectados, se adoptaron las siguientes estrategias

1.Eliminación de Outliers Extremos:

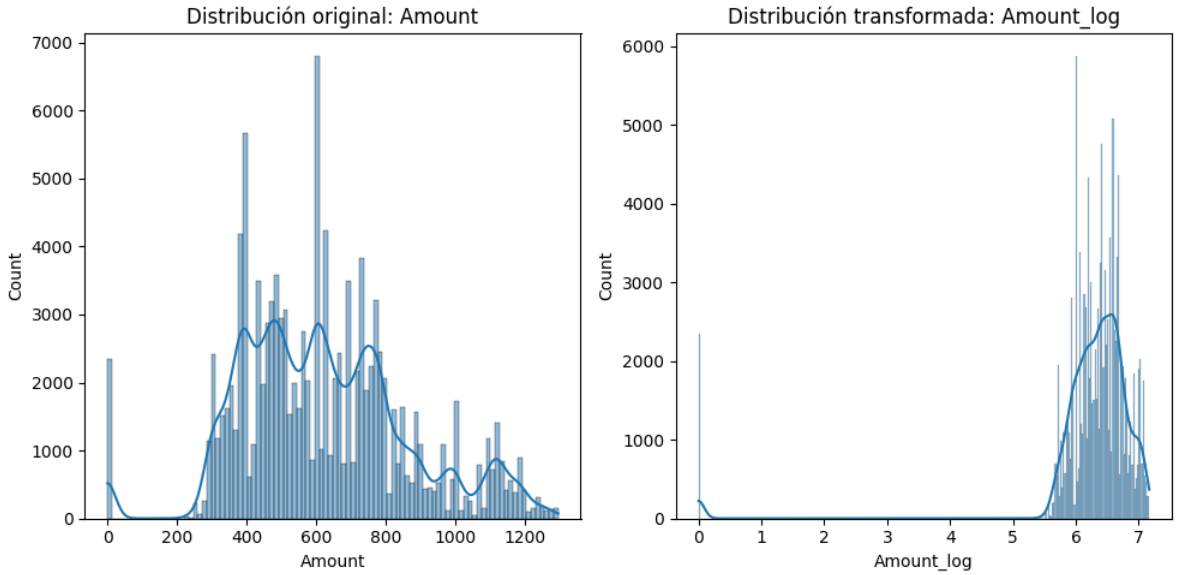
Se eliminaron valores que claramente representaban errores de entrada o eventos muy infrecuentes que no contribuyen al análisis general.

- **Qty:** Pedidos con cantidades superiores a 10 unidades fueron eliminados por ser casos poco representativos.
- **Amount:** Pedidos con montos superiores a 5,000 unidades monetarias fueron eliminados, excepto aquellos confirmados como válidos.
- **ship-postal-code:** Códigos postales que se encontraban fuera de los límites esperados o que correspondían a zonas no válidas fueron eliminados tras su verificación.

2.Transformación de Datos:

Para manejar distribuciones sesgadas y reducir el impacto de los valores atípicos:

- **Amount:** Se aplicó una transformación logarítmica para normalizar la distribución y minimizar el efecto de los valores extremos.

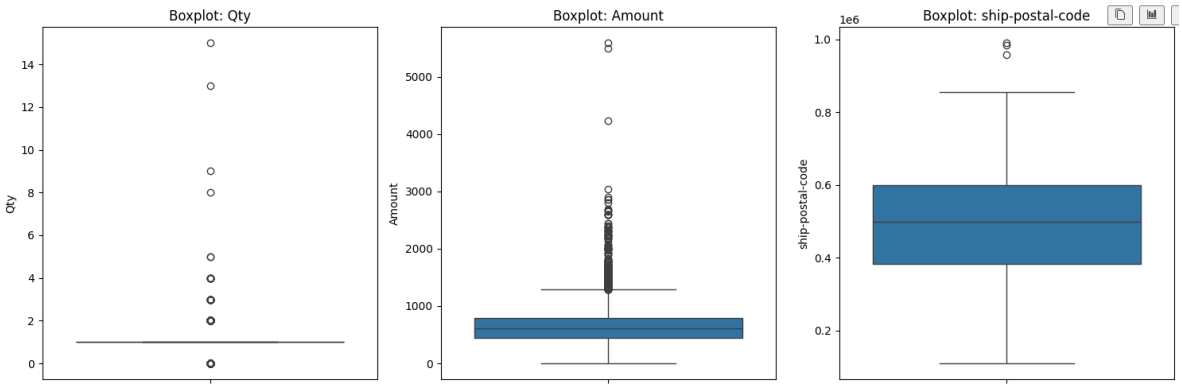


3.Imputación de Outliers:

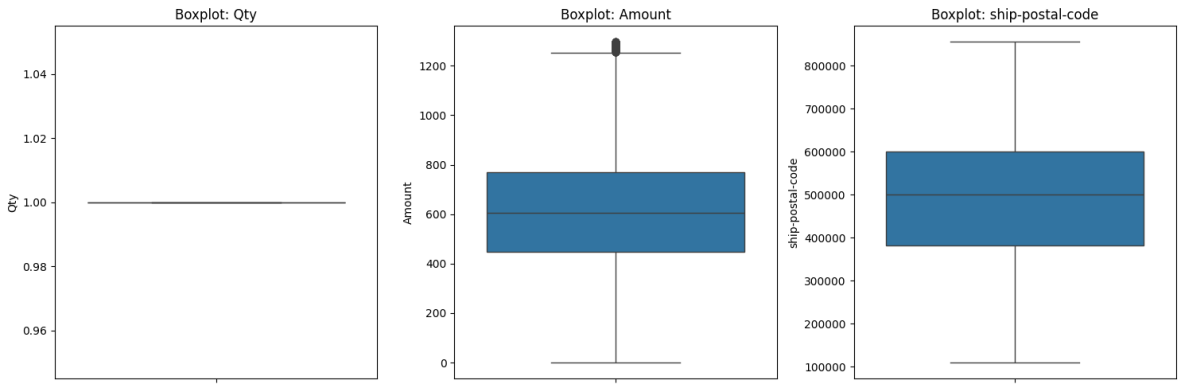
En los casos donde los valores atípicos representaban eventos válidos pero infrecuentes:

- **Qty:** Las cantidades altas fueron reemplazadas con la mediana de la variable para preservar la representatividad sin distorsionar los datos.
- **Amount:** Los montos elevados se imputaron con el valor promedio o la mediana según el contexto.
- **ship-postal-code:** Se imputaron valores faltantes o extremos con el código postal más frecuente en el área correspondiente.

Boxplot iniciales:



Boxplots después del tratamiento:



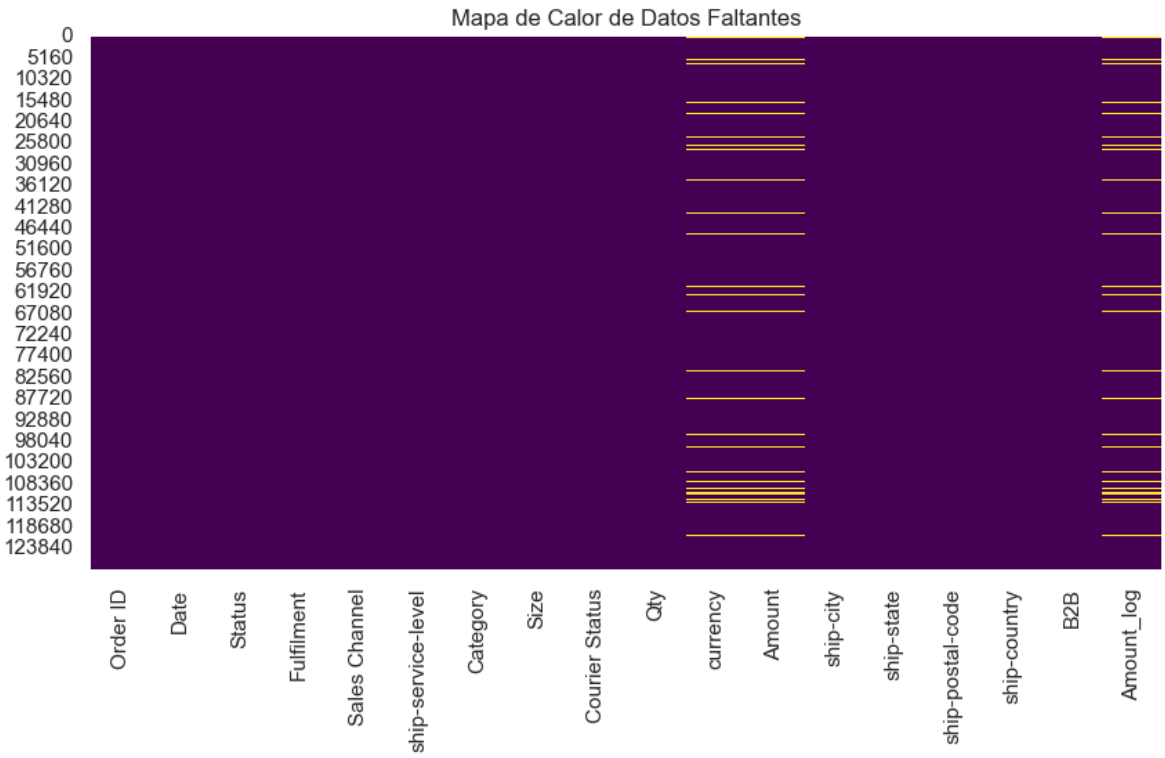
Justificación

Las decisiones sobre el tratamiento de los valores atípicos se realizaron considerando el impacto en los modelos predictivos y la calidad de los datos.

- **Eliminación:** Se usó para valores extremos que no aportan al análisis y podrían sesgar los resultados.
- **Transformación:** Se aplicó para reducir el efecto de sesgos en la distribución, lo que mejora el rendimiento del modelo.
- **Imputación:** Se utilizó para preservar información importante sin eliminar datos válidos.

5. Análisis de Valores Faltantes

Visualización de Datos Faltantes:



Identificación de Datos Faltantes:

Columna	Valores Faltantes	Porcentaje (%)
Order ID	**0**	**0.00%**
Date	**0**	**0.00%**
Status	**0**	**0.00%**
Fulfilment	**0**	**0.00%**
Sales Channel	**0**	**0.00%**
Ship-service-level	**0**	**0.00%**
Category	**0**	**0.00%**
Size	**0**	**0.00%**
Courier Status	**0**	**0.00%**
Qty	**0**	**0.00%**
Currency	**7800**	**6.05%**
Amount	**7800**	**6.05%**
Ship-city	**0**	**0.00%**
Ship-state	**0**	**0.00%**

Ship-postal-code	**35**	**0.03%**
Ship-country	**0**	**0.00%**
B2B	**0**	**0.00%**

Estrategia de Imputación o Eliminación:

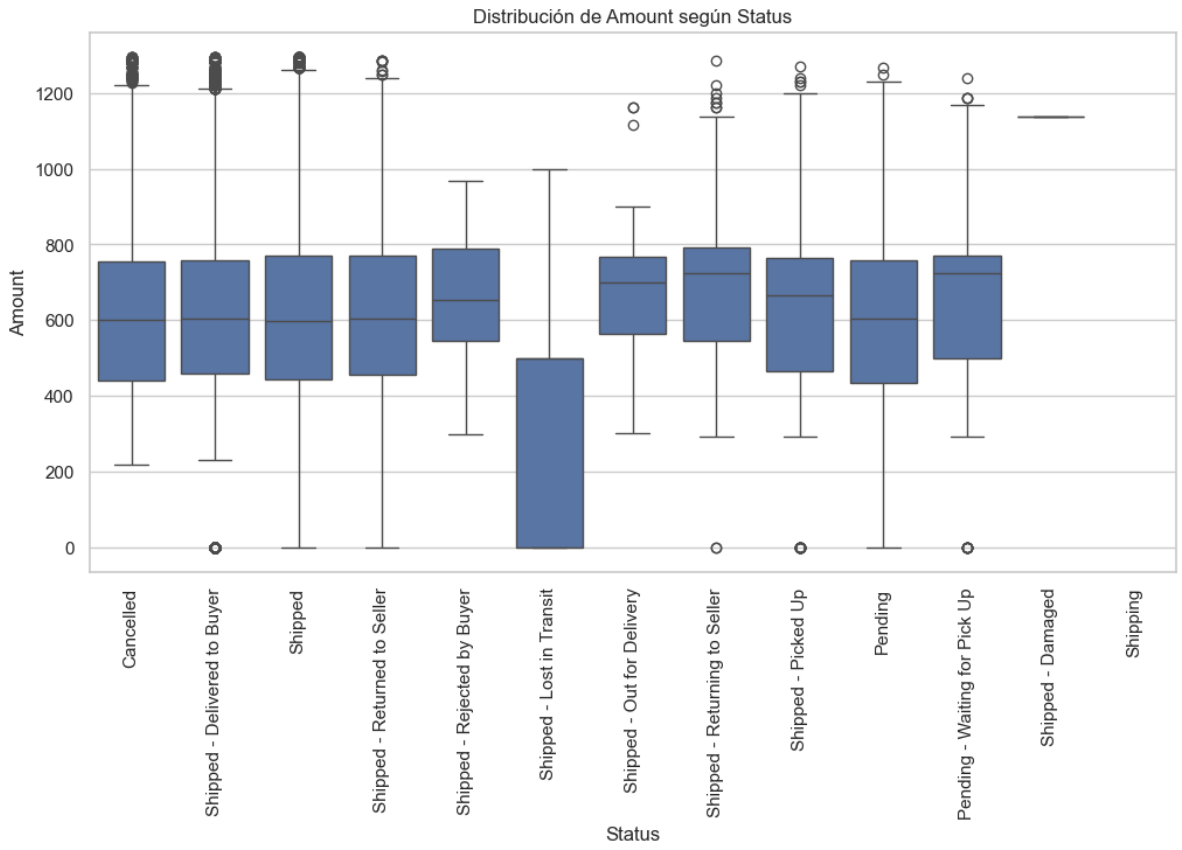
La columna **Amount** tienen un 6.05% de valores faltantes. Esto es significativo, pero no es un porcentaje lo suficientemente alto como para eliminar las columnas. En lugar de eliminarlas, es preferible imputar los valores faltantes.

Lo mismo pasa con la columna **Ship-postal-code** ya que tiene un 0.003% En lugar de eliminarlas, es preferible imputar los valores faltantes.

Dado que la columna **Currency** es categórica, la forma más adecuada de imputar los valores faltantes es utilizando la moda, es decir, el valor más frecuente de esa.

6. Relación entre Variables Categóricas y Numéricas

1. **Amount** según **Status** (Estado del Pedido):



La gráfica muestra la distribución de montos (Amount) según el estado (Status) de los pedidos.

Cancelled:

- Montos: Generalmente bajos, con la mayoría de los valores cercanos a 0.
- Interpretación: Los pedidos cancelados tienden a tener montos pequeños, lo que podría indicar que los pedidos de mayor valor son menos propensos a ser cancelados.

Shipped:

- Montos: Varían ampliamente, con valores que van desde bajos hasta altos.
- Interpretación: Este estado muestra una gran variabilidad en los montos, sugiriendo que los pedidos enviados pueden ser de cualquier valor.

Shipped - Returned to Seller:

- Montos: Predominantemente bajos, con algunos valores más altos.

- Interpretación: Los pedidos devueltos al vendedor suelen tener montos más bajos, aunque hay excepciones.

Shipped - Rejected by Buyer:

- Montos: Varían ampliamente, con algunos valores altos.
- Interpretación: Los pedidos rechazados por el comprador muestran una gran variabilidad en los montos, indicando que pueden ser de cualquier valor.

Shipped - Lost in Transit:

- Montos: Generalmente bajos, con algunos valores medianos.
- Interpretación: Los pedidos perdidos en tránsito tienden a tener montos menores, lo que podría indicar que los pedidos de mayor valor tienen un seguimiento más riguroso.

Shipped - Out for Delivery:

- Montos: Varían ampliamente, con valores que van desde bajos hasta altos.
- Interpretación: Este estado muestra una gran variabilidad en los montos, sugiriendo que los pedidos en entrega pueden ser de cualquier valor.

Shipped - Returning to Seller:

- Montos: Predominantemente bajos, con algunos valores más altos.
- Interpretación: Los pedidos que están siendo devueltos al vendedor suelen tener montos más bajos, aunque hay excepciones.

Pending:

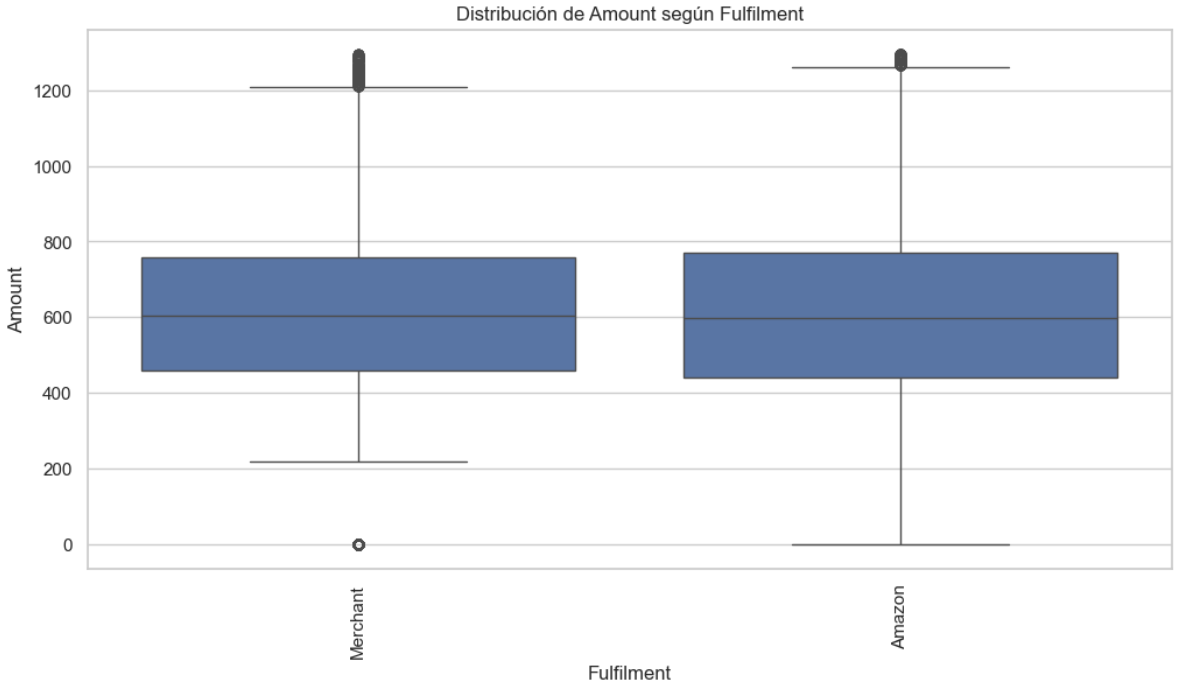
- Montos: Generalmente bajos, con algunos valores medianos.
- Interpretación: Los pedidos pendientes tienden a tener montos menores, lo que podría indicar que los pedidos de mayor valor se procesan más rápidamente.

Tendencias Generales:

- Pedidos Cancelados y Pendientes: Tienden a tener montos más bajos.

- Pedidos Enviados y Entregados: Muestran una mayor variabilidad en los montos, incluyendo valores altos.
- Pedidos Devueltos: Aunque generalmente tienen montos bajos, pueden incluir algunos pedidos de mayor valor.

2. **Amount** según **Fulfilment** (Método de Cumplimiento):

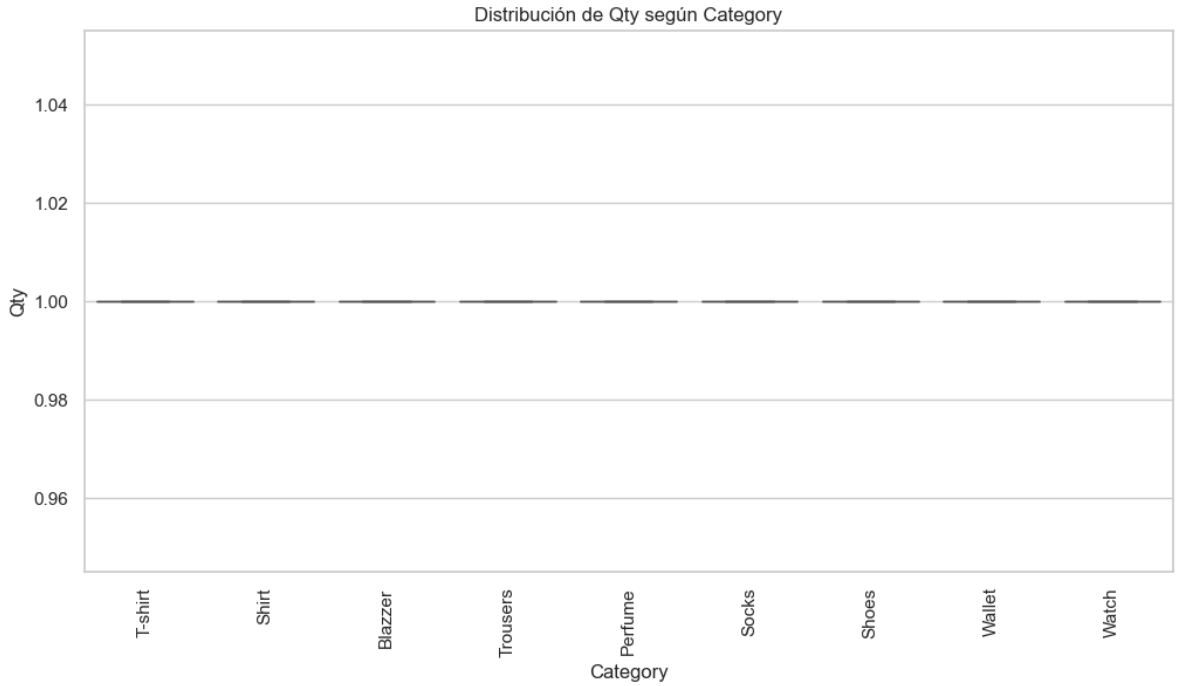


La grafica muestra la Distribución de (Amount) según (Fulfilment)

- Los pedidos cumplidos directamente por Amazon (Amazon) tienen montos más altos en promedio que los cumplidos por vendedores externos (Merchant).
- Amazon tiene una menor variabilidad en comparación con (Merchant)
- Ambas categorías tienen outliers, lo que indica la presencia de valores atípicos.

Esto podría reflejar una mayor confianza de los clientes en los productos manejados directamente por Amazon.

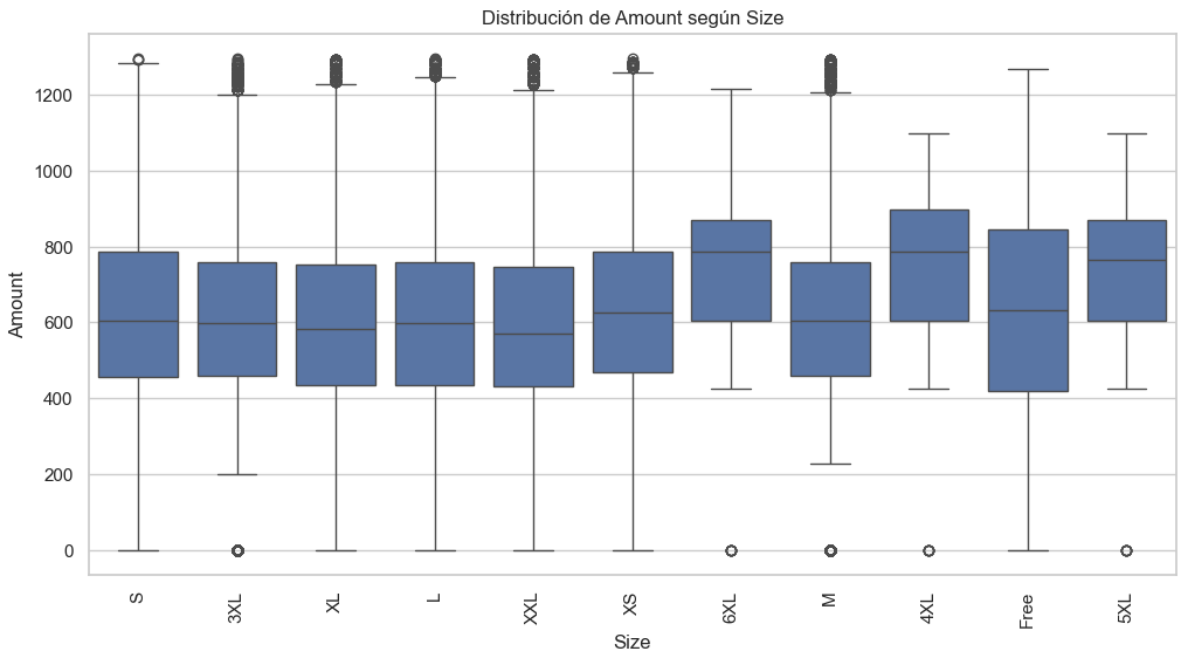
3. Qty según **Category** (Categoría del Producto):



Este gráfico boxplot muestra una uniformidad en las cantidades (Qty) compradas para cada categoría (Category) de producto, con un valor constante de 1

- Sin Variabilidad: No hay variabilidad ni outliers en las cantidades, lo que indica que regularmente se compra una cantidad fija por categoría.

4. **Amount** según **Category** (Categoría del Producto):

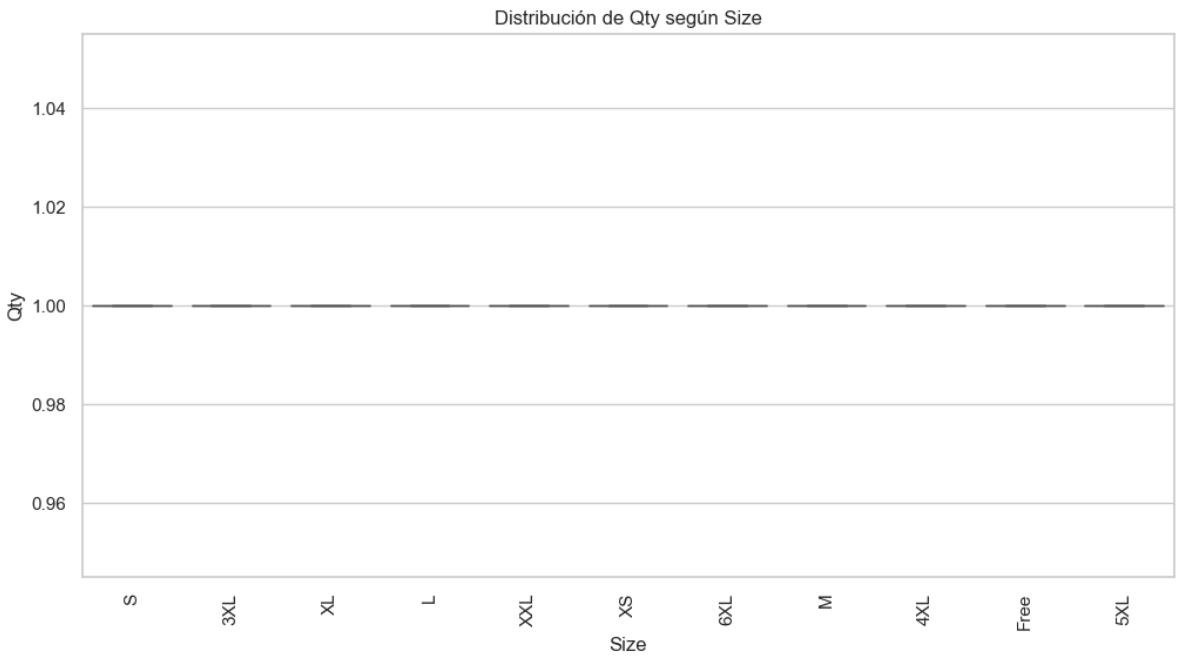


La gráfica muestra la distribución de montos (Amount) según el tamaño (Size) de los productos.

Tendencias Generales:

- Las categorías de tamaño como ‘XL’ muestran una mayor variabilidad en la cantidad, con un rango intercuartil más amplio en comparación con tamaños como ‘S’.
- Los tamaños con una mayor dispersión en los boxplots tienen una mayor variabilidad en los montos.

5. **Qty** según **Size** (Talla del Producto):



Dado que solo se compra regularmente una cantidad por cada tamaño de prenda, el boxplot refleja esta uniformidad en las cantidades.

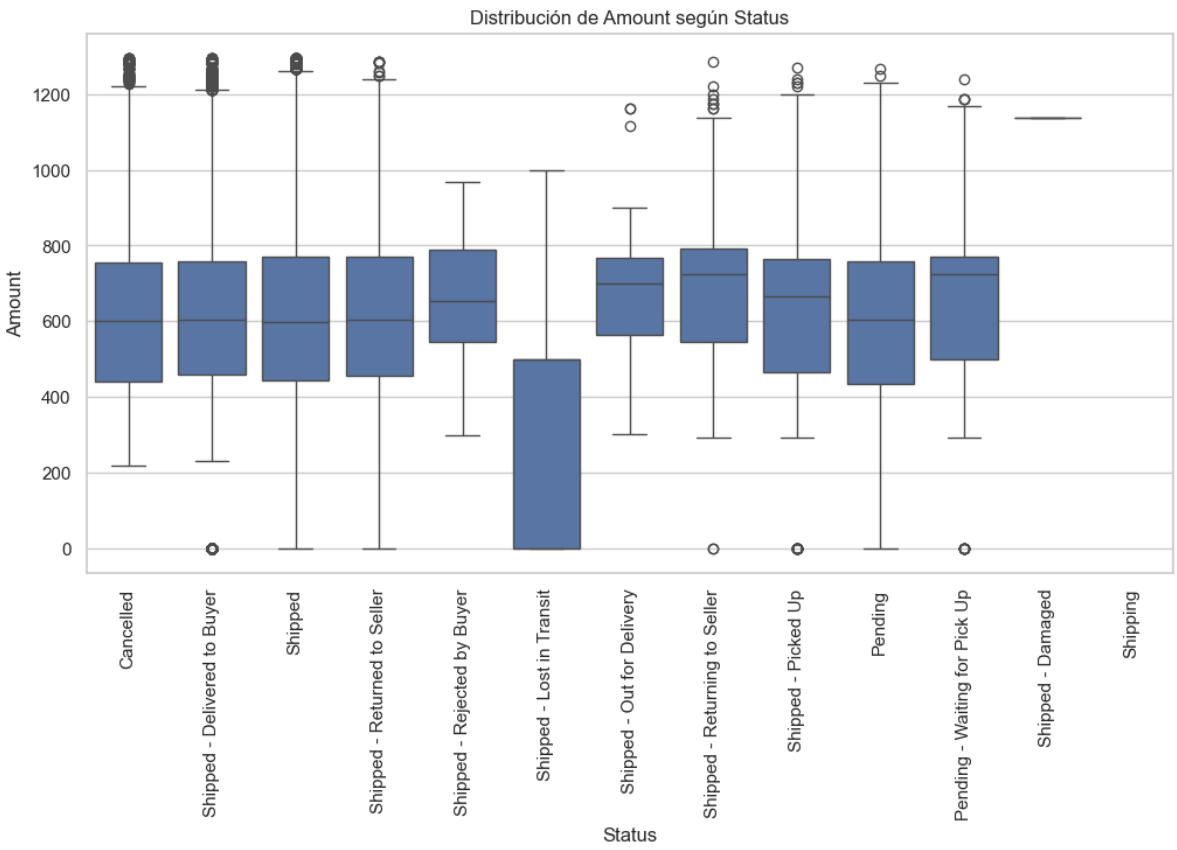
Tamaños

- Montos: Cada tamaño tiene una cantidad de 1, lo que resulta en una distribución uniforme.
- Interpretación: La falta de variabilidad en las cantidades hace que los boxplots sean líneas horizontales en el valor de 1, sin mostrar dispersión ni outliers.

Tendencias Generales:

- Uniformidad: La compra de una sola cantidad por cada tamaño elimina cualquier variabilidad en los datos, resultando en boxplots que no muestran diferencias entre categorías.
- Visualización: Aunque los boxplots son útiles para comparar distribuciones, en este caso específico, no aportan información adicional debido a la uniformidad de los datos.

6. **Amount** según **Courier Status** (Estado del Mensajero):



La gráfica muestra la distribución de montos (Amount) según el estado (Status) de los pedidos.

Cancelled:

- Montos: Generalmente bajos, con la mayoría de los valores cercanos a 0.
- Interpretación: Los pedidos cancelados tienden a tener montos pequeños, lo que podría indicar que los pedidos de mayor valor son menos propensos a ser cancelados.

Shipped - Delivered to Buyer:

- Montos: Varían ampliamente, con valores que van desde bajos hasta altos.
- Interpretación: Este estado muestra una gran variabilidad en los montos, sugiriendo que los pedidos entregados pueden ser de cualquier valor.

Shipped - Returned to Seller:

- Montos: Predominantemente bajos, con algunos valores más altos.
- Interpretación: Los pedidos devueltos al vendedor suelen tener montos más bajos, aunque hay excepciones.

Pending:

- Montos: Generalmente bajos, con algunos valores medianos.
- Interpretación: Los pedidos pendientes tienden a tener montos menores, lo que podría indicar que los pedidos de mayor valor se procesan más rápidamente.

Tendencias Generales:

- Pedidos Cancelados y Pendientes: Tienden a tener montos más bajos.
- Pedidos Enviados y Entregados: Muestran una mayor variabilidad en los montos, incluyendo valores altos.
- Pedidos Devueltos: Aunque generalmente tienen montos bajos, pueden incluir algunos pedidos de mayor valor.

7. Observaciones y Hallazgos Importantes

Declaración de la Variable Objetivo y Variables Relacionadas

Variable objetivo: Estado del pedido (Status), con la meta de reducir la tasa de cancelación de pedidos (Status = "Cancelled").

Variables relacionadas: Se observan varias variables que afectan el estado del pedido, tales como:

- Cantidad de productos (**Qty**), donde se observa que los pedidos con mayor cantidad de productos tienen menos probabilidades de ser cancelados.
- Monto total del pedido (**Amount**), donde los pedidos con montos más altos tienden a ser menos cancelados, lo que puede indicar que los clientes son menos propensos a cancelar pedidos grandes.
- Método de cumplimiento (**Fulfilment**) y Estado del mensajero (Courier Status), que afectan la logística y el tiempo de entrega, siendo variables

clave para la cancelación de pedidos debido a demoras o problemas con el envío.

- Categoría de producto (Category), con ciertas categorías que muestran una mayor tendencia a la cancelación.

Resumen de Hallazgos Clave

Relaciones inesperadas:

- Cantidad de productos (Qty) muestra una relación débil con la cancelación de pedidos, lo que podría sugerir que los pedidos con una sola unidad no son más propensos a ser cancelados que los de mayor cantidad.
- Monto total (Amount) tiene una correlación moderada con la cancelación. Sin embargo, los montos más altos no siempre protegen contra la cancelación, lo que podría implicar factores adicionales como la disponibilidad de productos.

Patrones interesantes:

- Nivel de servicio de envío (ship-service-level) tiene una fuerte relación con el Método de cumplimiento (Fulfilment), lo que podría significar que optimizar el cumplimiento de los pedidos a través de Amazon podría mejorar las tasas de entrega y reducir las cancelaciones.
- Categorías de productos más caras y populares tienden a tener una tasa de cancelación más baja, lo que puede estar relacionado con una mayor confianza en el vendedor o en la marca.

Variables con fuertes correlaciones:

- La fuerte correlación entre Método de cumplimiento (Fulfilment) y Nivel de servicio de envío (ship-service-level) (0.98) sugiere que ambas variables podrían ser redundantes en el modelo por lo que es mejor la eliminación de una.

- La relación entre Estado del mensajero (Courier Status) y Cantidad (Qty) (0.83) sugiere que el estado de envío podría ser predicho por la cantidad de productos en el pedido, ayudando a identificar pedidos que están cerca de la entrega o en camino.
- Se observó que ciertos productos, como T-shirts y Shirts, tienen una mayor cantidad de unidades por pedido, lo cual podría indicar que estos productos se compran en lotes, lo que a su vez podría influir en las cancelaciones debido a errores en el inventario o el proceso logístico.

Anomalías:

- Se observó que ciertos productos, como T-shirts y Shirts, tienen una mayor cantidad de unidades por pedido, lo cual podría indicar que estos productos se compran en lotes, lo que a su vez podría influir en las cancelaciones debido a errores en el inventario o el proceso logístico.

Implicaciones para el Modelo

Selección de variables: Las variables más relevantes para predecir la cancelación de pedidos son Monto total (Amount), Cantidad de productos (Qty), Método de cumplimiento (Fulfilment), Estado del mensajero (Courier Status), y Categoría del producto (Category).

Interpretación de resultados: El modelo debe considerar que los pedidos con un mayor número de productos y un monto más alto tienen menos probabilidades de ser cancelados, lo que sugiere que la optimización debe centrarse en mejorar la experiencia del cliente en relación con el cumplimiento, el envío y la disponibilidad de productos.

4. Modelo de Machine Learning

Descripción del modelo: Se utilizó un árbol de decisión (DecisionTreeClassifier) para predecir la cancelación de pedidos en Amazon. Este modelo es apropiado

para problemas de clasificación y permite dividir los datos en "nodos" para hacer predicciones basadas en decisiones binarias.

Justificación: El árbol de decisión fue elegido debido a su capacidad para manejar tanto variables numéricas como categóricas, su interpretación visual fácil y la simplicidad en la implementación. Además, al ser un modelo no lineal, es capaz de capturar relaciones complejas entre las características del dataset.

Implementación y Entrenamiento:

- **División de datos:** Se separaron los datos en dos conjuntos: un conjunto de entrenamiento (70%) y uno de prueba (30%).
- **Métricas:** Se utilizó la precisión (accuracy) y el informe de clasificación (classification report), que incluye las métricas de precisión, recall y F1-score.
- **Ajustes de parámetros:** Se ajustaron los parámetros del árbol de decisión, tales como la profundidad máxima (max_depth=10) y el número mínimo de muestras por hoja (min_samples_leaf=5) para evitar el sobreajuste y mejorar la generalización del modelo.

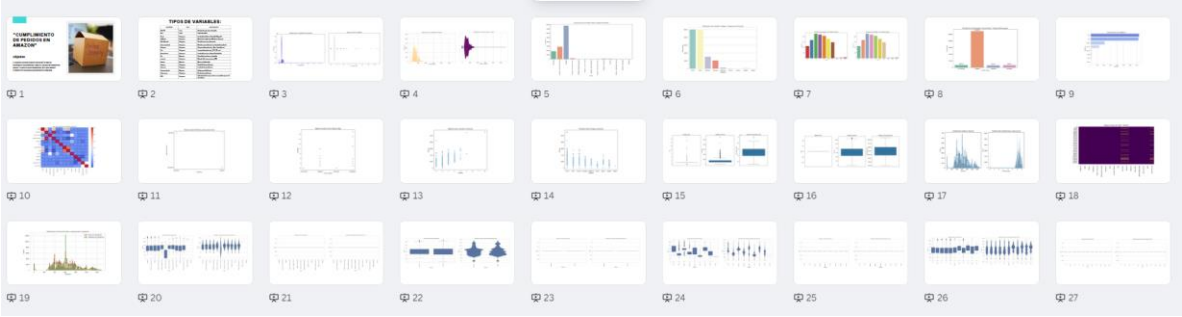
Resultados:

```
Accuracy: 0.9992026347720575
Classification Report:
               precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     23283
         1.0       1.00      1.00      1.00      5562

    accuracy                           1.00     28845
   macro avg       1.00      1.00      1.00     28845
weighted avg       1.00      1.00      1.00     28845
```

El modelo presentó un accuracy de 0.9992, lo que indica una precisión casi perfecta en la clasificación de los pedidos. El informe de clasificación muestra que tanto la precisión como el recall son 1.00 para ambas clases (cancelados y no cancelados).

5. Dashboard

Capturas y Explicación del Dashboard:



El objetivo del dashboard es proporcionar una visualización clara y accesible de los datos clave relacionados con el cumplimiento de pedidos y la tasa de cancelación.

Uso y Beneficios:

El uso del dashboard es esencial para la toma de decisiones basada en datos.

Algunos de los principales beneficios son:

Análisis centrado en áreas críticas: Ayudar a identificar áreas críticas donde se pueden implementar mejoras para reducir la tasa de cancelación y optimizar el cumplimiento de pedidos.

Toma de decisiones informadas: proporcionará información útil para prever tendencias y tomar decisiones informadas para mejorar los procesos, ajustar inventarios o mejorar la satisfacción del cliente.

6. Conclusiones y Futuras Líneas de Trabajo

Resumen de los hallazgos principales y cómo estos cumplen con los objetivos planteados al inicio:

A lo largo de este proyecto, se ha logrado analizar la tasa de cancelación de pedidos en Amazon y se han identificado diversos factores que afectan este fenómeno, tales como la cantidad de productos en un pedido y las categorías de productos.

A través de la implementación de un modelo de machine learning, específicamente un árbol de decisión, se ha logrado predecir con una precisión de

(99.92%) si un pedido será cancelado o no, lo que ha proporcionado una base sólida para comprender los factores que influyen en las cancelaciones.

Los insights obtenidos, como la tendencia de cancelaciones por categoría de productos permiten a Amazon identificar áreas clave en las que se pueden implementar estrategias como justar el inventario de productos con mayores tasas de cancelación, para mitigar esta misma cancelación de pedidos.

Posibles mejoras:

- **Ampliar el Dataset:** El modelo se basó en un conjunto de datos relativamente limitado. Para mejorar la generalización del modelo, sería beneficioso incorporar más datos, especialmente en lo que respecta a variables adicionales, como el tiempo de envío o la comunicación con el cliente, que podrían tener un impacto significativo en las cancelaciones.
- **Incluir variables de comportamiento del cliente:** Incorporar variables relacionadas con el comportamiento del cliente (por ejemplo, historial de compras, frecuencia de devoluciones, etc.) podría ayudar a afinar aún más las predicciones.

Posibles Direcciones para Investigaciones Futuras:

Análisis de Satisfacción del Cliente: Realizar un análisis de cómo las cancelaciones afectan la satisfacción del cliente y el comportamiento futuro de compra.

7. Referencias

Bases de Datos:

- Dataset utilizado: *Base_limpia.csv*, proporcionado desde el repositorio: https://raw.githubusercontent.com/Jylians/introduc/refs/heads/main/Base_limpia.csv

Documentación Técnica y Bibliografía:

- Scikit-learn: Herramienta utilizada para el entrenamiento y evaluación del modelo de machine learning.
  Página oficial: https://scikit-learn.org/stable/

- Pandas: Librería utilizada para la manipulación y análisis de datos.
  Página oficial: https://pandas.pydata.org/docs/
- Matplotlib: Librería utilizada para la creación de visualizaciones básicas en Python.
  Página oficial: https://matplotlib.org/stable/contents.html
- Seaborn: Herramienta utilizada para mejorar visualizaciones de datos en Python.
  Página oficial: https://seaborn.pydata.org/

Recursos en Línea:

- Kaggle: Your First Machine Learning Model- Referencias para la evaluación de modelos y manejo de datos.
  URL: https://www.kaggle.com/code/dansbecker/your-first-machine-learning-model

8. Anexos

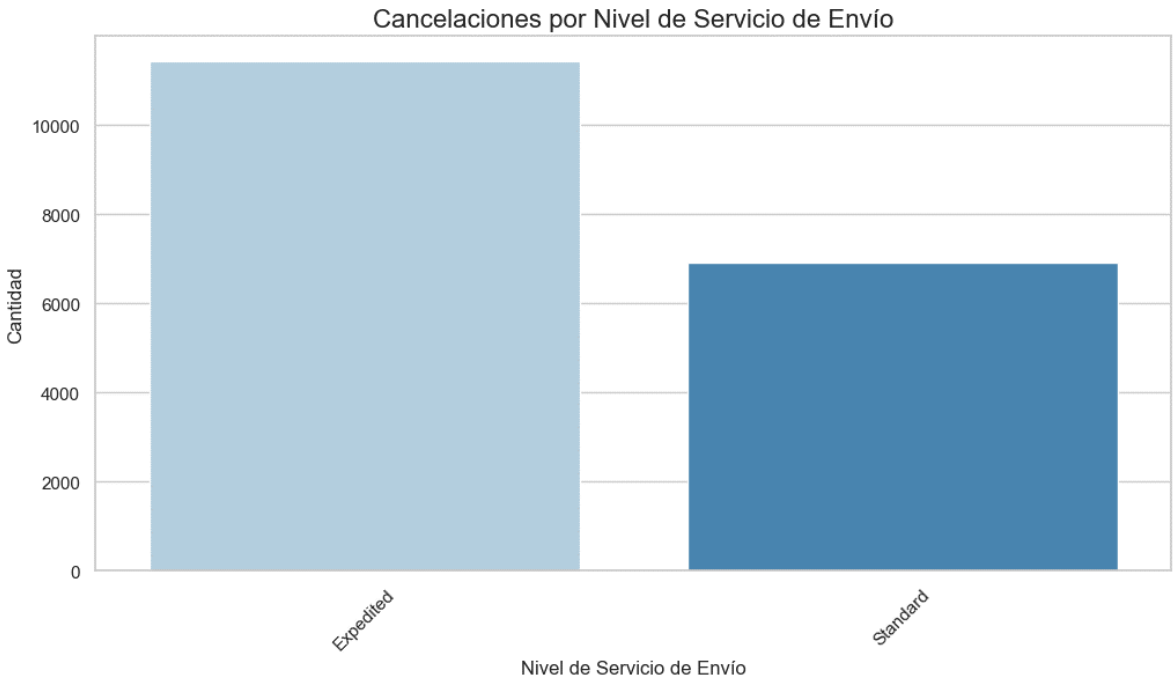
A) Código Fuente Relevante

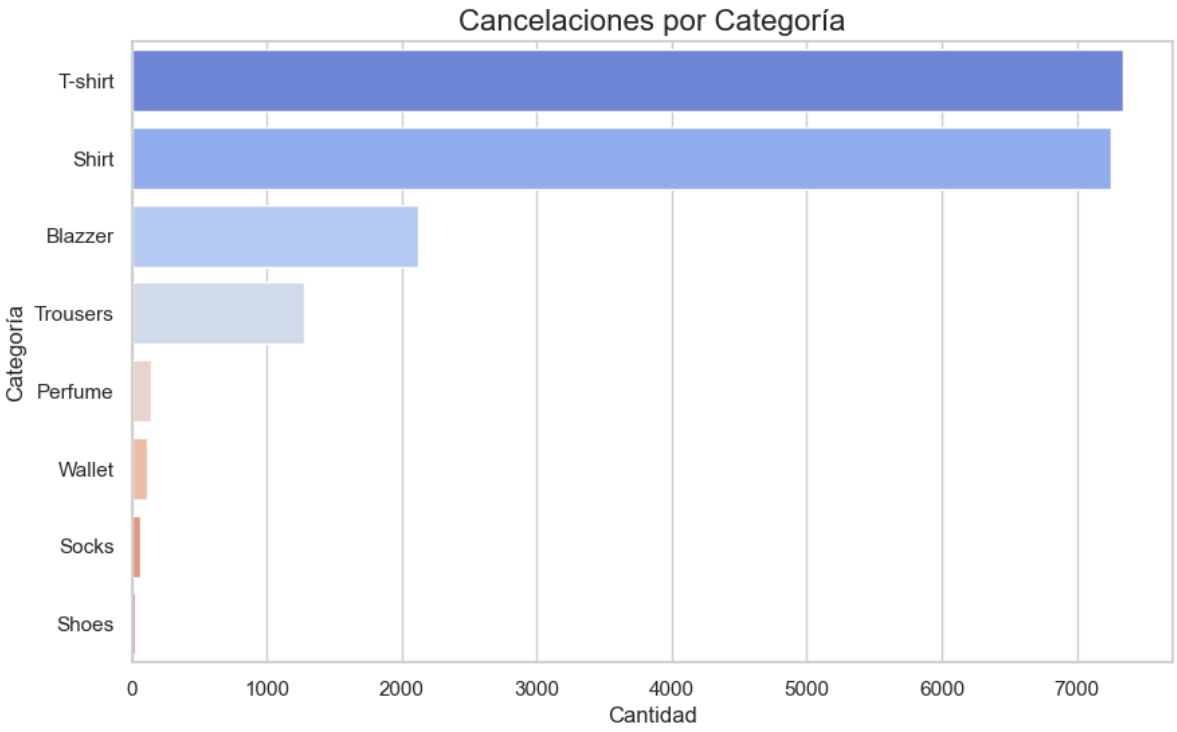
El código utilizado para la implementación del modelo de machine learning.

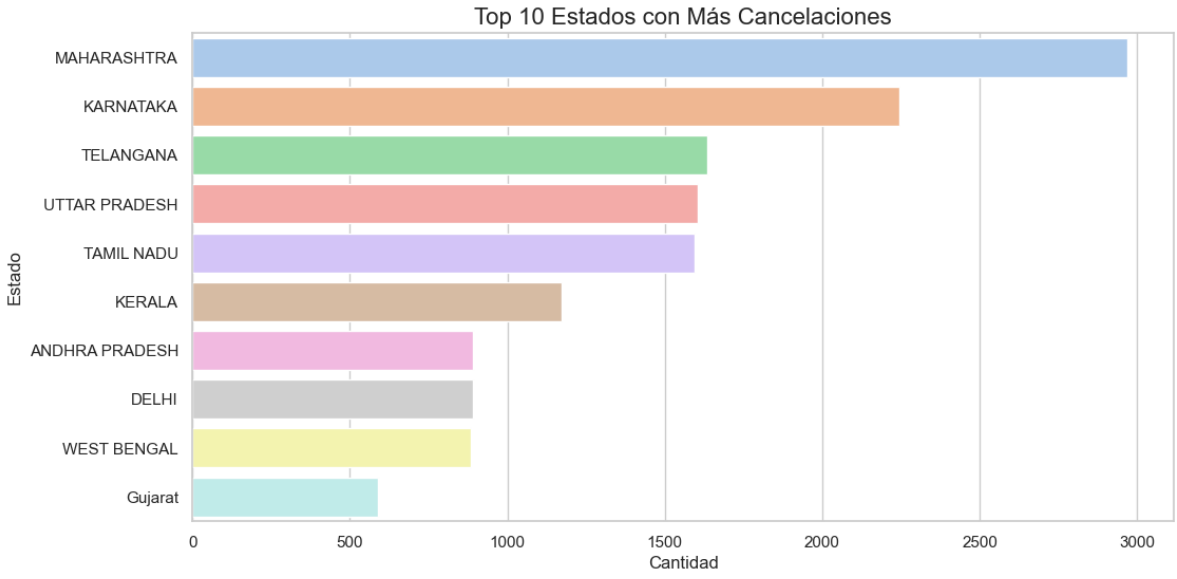
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report, accuracy_score


# Cargar el dataset
df=pd.read_csv("https://raw.githubusercontent.com/Jylians/introduc/refs/heads/main/Base_limpia.csv")


# Paso 1: Preprocesamiento
```

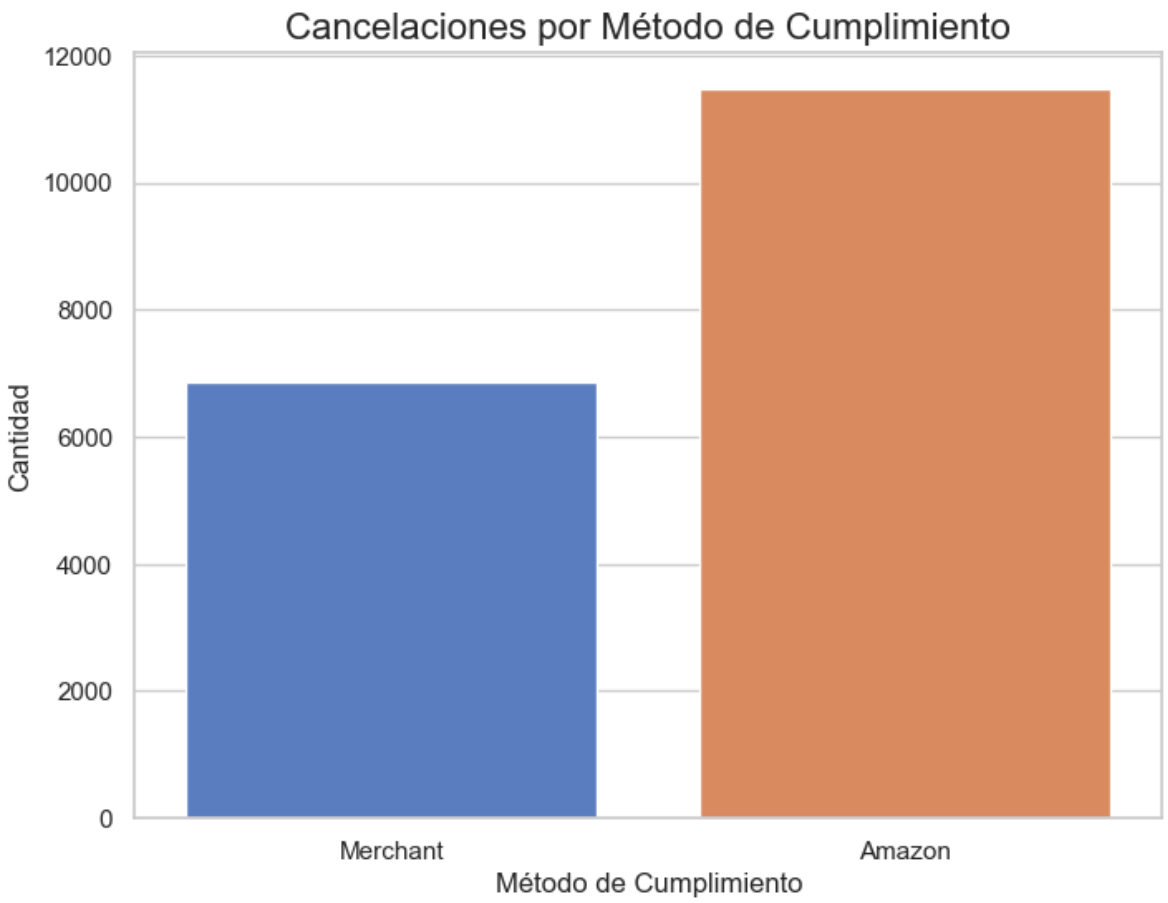
```
# Manejo de valores nulos
df['Amount'].fillna(df['Amount'].median(), inplace=True)
df['Qty'].fillna(df['Qty'].median(), inplace=True)

# Codificación de variables categóricas
label_encoder = LabelEncoder()
df['Fulfilment'] = label_encoder.fit_transform(df['Fulfilment'])
df['Category'] = label_encoder.fit_transform(df['Category'])
df['Courier Status'] = label_encoder.fit_transform(df['Courier Status'])
df['currency'] = label_encoder.fit_transform(df['currency'])

# Limpieza de la columna 'Status'
df['Status'] = df['Status'].map({'Cancelled': 1, 'Shipped': 0})  # Cancelado = 1, No Cancelado = 0
df.dropna(subset=['Status'], inplace=True)

# Paso 2: División de datos
X = df[['Amount', 'Qty', 'Fulfilment', 'Category', 'Courier Status']]
y = df['Status']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Paso 3: Entrenamiento del modelo
model = DecisionTreeClassifier(max_depth=10, min_samples_leaf=5, random_state=42)
model.fit(X_train, y_train)

# Evaluación del modelo
y_pred = model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Classification Report:\n", classification_report(y_test, y_pred))
```

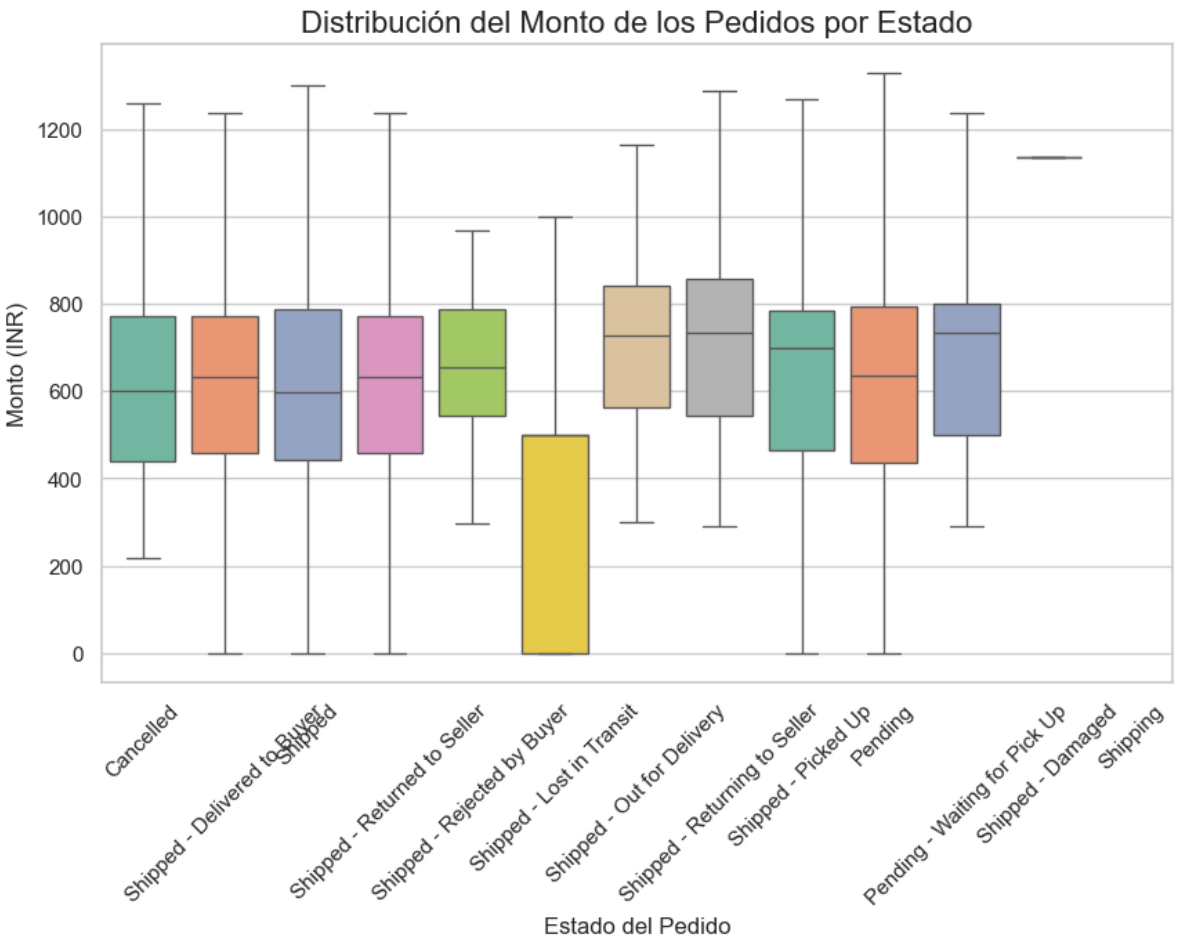
B) Gráficos o análisis adicionales no incluidos en el reporte principal.

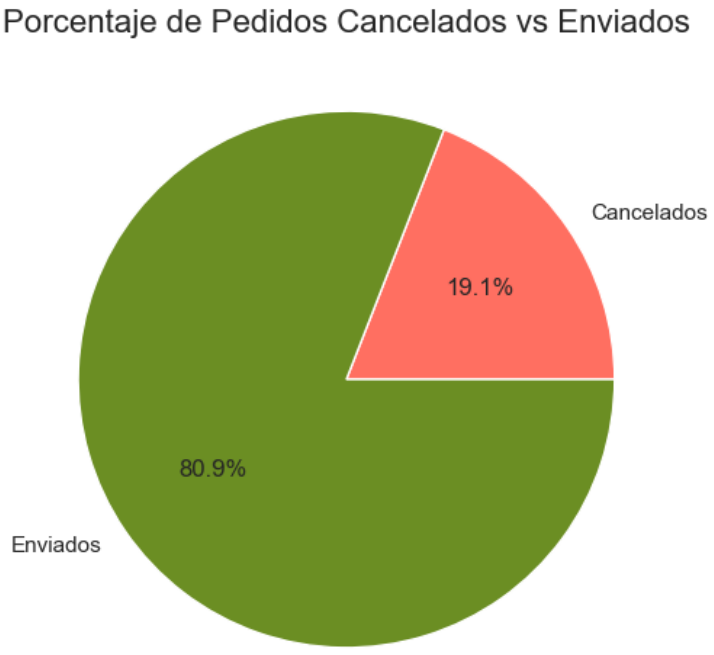



Cancelaciones por Categoría
T-shirt
Shirt
Blazzer
Trousers
Perfume
Wallet
Socks
Shoes
Categoría
0
1000
2000
3000
4000
5000
6000
7000
Cantidad


Top 10 Estados con Más Cancelaciones
MAHARASHTRA
KARNATAKA
TELANGANA
UTTAR PRADESH
TAMIL NADU
KERALA
ANDHRA PRADESH
DELHI
WEST BENGAL
Gujarat
Estado
0
500
1000
1500
2000
2500
3000
Cantidad


Cancelaciones por Método de Cumplimiento
12000
10000
8000
6000
4000
2000
0
Cantidad
Merchant
Amazon
Método de Cumplimiento


Distribución del Monto de los Pedidos por Estado
Monto (INR)
0
200
400
600
800
1000
1200
Cancelled
Shipped - Delivered to Buyer
Shipped
Shipped - Returned to Seller
Shipped - Rejected by Buyer
Shipped - Lost in Transit
Shipped - Out for Delivery
Shipped - Returning to Seller
Shipped - Picked Up
Pending
Pending - Waiting for Pick Up
Shipped - Damaged
Shipping
Estado del Pedido


Porcentaje de Pedidos Cancelados vs Enviados
Cancelados
19.1%
80.9%
Enviados

Base de datos limpia que se utilizó

https://raw.githubusercontent.com/Jylians/introduc/refs/heads/main/Base_limpia.csv