# Proyecto Módulo 4: Análisis Exploratorio de Datos para decisiones comerciales

Nombre: Luis León Barrera

Curso: Fundamentos de Ciencia de Datos

Fecha: 18/02/2026

# 1. Introducción

Esta es una empresa dedicada a la venta de productos automotrices, que ha acumulado un registro histórico de ventas con información de productos, documentos tributarios, montos netos, impuestos, costos y ganancias. El equipo directivo requiere una visión estructurada de estos datos para comprender mejor el comportamiento de las ventas y tomar decisiones comerciales informadas.

El análisis se realizó en Python utilizando pandas, seaborn, matplotlib y statsmodels, sobre un dataset de 1450 registros de la hoja “Ventas” del archivo de venta del comercio.

# 2. Descripción del dataset y preparación de datos

El dataset original contiene, entre otras, las siguientes variables principales:

- **Producto:** descripción del producto o paquete vendido.
- **Fecha:** fecha de la transacción.
- **DTE:** tipo de documento (Boleta, Factura, C. Venta).
- **N. DTE:** número de documento tributario.
- **Neto, IVA, Total:** montos netos, impuesto y total de la venta.
- **Costo venta:** costo asociado a la venta.
- **Ganancia Bruta:** margen bruto registrado.

## 2.1. Limpieza de datos

Las principales acciones de limpieza fueron:

- Conversión de **Fecha** a tipo datetime y corrección de dos valores faltantes mediante forward fill (copiando la fecha de la fila anterior, dado que el dataset está ordenado cronológicamente).
- Conversión de **N. DTE**, **Costo venta**, **Total** y **Ganancia Bruta** a tipo numérico, resolviendo valores no numéricos como faltantes.
- Relleno de **Neto** e **IVA** cuando faltaban

## 2.2. Enriquecimiento de datos

Con el fin de enriquecer el análisis, se incorporaron columnas derivadas y simuladas de forma controlada:

- **Medio_de_pago:** asignado aleatoriamente entre Transferencia, Efectivo, Tarjeta Débito y Tarjeta Crédito, con probabilidades definidas para reflejar un escenario realista.
- **estado_pago:** estado de pago (Pagado, Pendiente, Parcial) condicionado al medio de pago (por ejemplo, mayor proporción de "Pagado" en tarjetas).
- **Tipo Cliente:**
    - Factura → Empresa (clientes B2B).
    - Boleta/C. Venta/sin DTE → asignación aleatoria entre Nuevo, Frecuente y VIP, con mayor peso en clientes frecuentes.
- **Entrega:** Local o Domicilio, simulando canal de entrega.
- **Categoría_venta:** Tradicional o Ecommerce, para diferenciar el canal de venta.
- **Dias_desde_compra:** diferencia en días entre la fecha de referencia (máxima fecha del dataset) y la fecha de la venta.
- **Cantidad:** número de ítems aproximado en la descripción del producto, contando segmentos separados por coma o "+" (por ejemplo, paquetes de filtros y aceites).
- **estacionalidad:** trimestre del año (Q1, Q2, Q3, Q4) según el mes de la venta.

Para variables financieras incompletas:

- **Ganancia Bruta:** cuando faltaba y existía un Total válido, se imputó asumiendo un margen bruto aleatorio entre 30% y 40% del Total, con fines exclusivamente analíticos.
- **Costo venta:** cuando era cero o nulo y existía información de Total y Ganancia Bruta, se calculó como:

  Costo venta = Total − Ganancia Bruta.

# 3. Análisis exploratorio de datos

## 3.1. Visión general y distribución de variables

- El dataset final incluye 1450 ventas con información completa de montos y fechas.
- Las variables numéricas analizadas incluyen: Neto, IVA, Total, Costo venta, Ganancia Bruta, Cantidad y Dias_desde_compra.
- La mayoría de las ventas se concentran en montos de Total entre aproximadamente 20.000 y 50.000, con algunos valores altos correspondientes a ventas de neumáticos, paquetes de mantención y servicios más complejos.
- La variable Cantidad muestra que la mayoría de las ventas corresponden a 1 o 2 ítems en la descripción, con algunos casos hasta 7 ítems.

En términos de tipos de variables:

- **Numéricas:** N. DTE, Neto, IVA, Total, Costo venta, Ganancia Bruta, Dias_desde_compra, Cantidad.
- **Categóricas:** DTE, Medio_de_pago, estado_pago, Tipo Cliente, Entrega, Categoría_venta, estacionalidad, Producto.

## 3.2. Medidas de tendencia central y dispersión

Sobre las principales variables financieras se calcularon media, mediana, varianza, desviación estándar, cuartiles y percentiles:

- **Total:**
    - Media cercana a 42.000; mediana alrededor de 33.000.
    - Desviación estándar alta, lo que confirma la presencia de ventas de mayor valor que elevan la dispersión.
- **Neto e IVA:**
    - Se mantienen coherentes con la tasa de IVA del 19%, con alta correlación entre ambas y con Total.
- **Costo venta y Ganancia Bruta:**
    - El costo suele situarse entre 20.000 y 35.000 para la mayoría de las ventas, con márgenes brutos típicos entre 30% y 40%, de acuerdo al supuesto de imputación.

- **Cantidad:**
    - Mediana = 1, lo que indica que la mayoría de las ventas concentran uno o pocos productos; los casos de Cantidad alta corresponden a paquetes más complejos.

Boxplots e histogramas permitieron identificar:

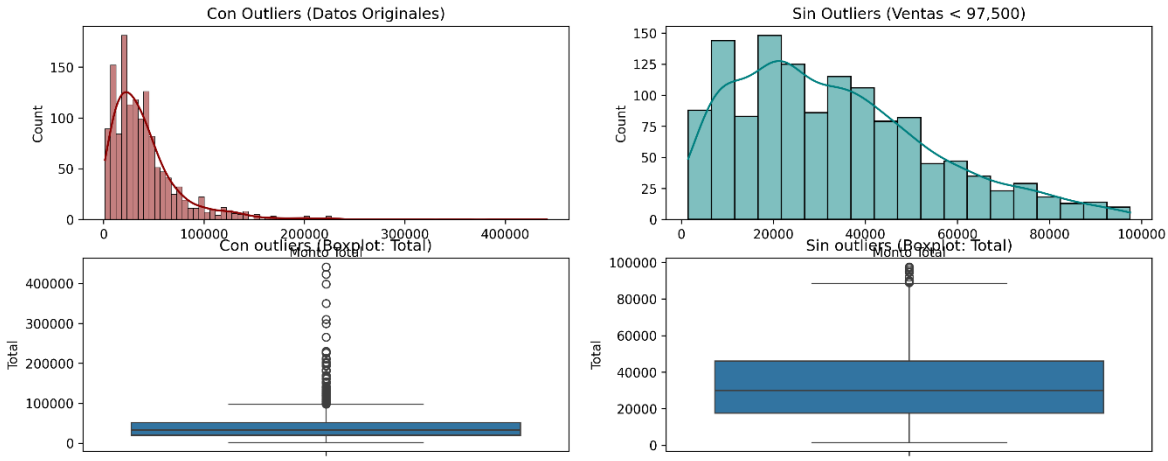
- Distribuciones sesgadas a la derecha en Total y Ganancia Bruta (colas largas hacia montos altos).
- Concentración de ventas en un rango relativamente acotado, con pocos casos extremos.

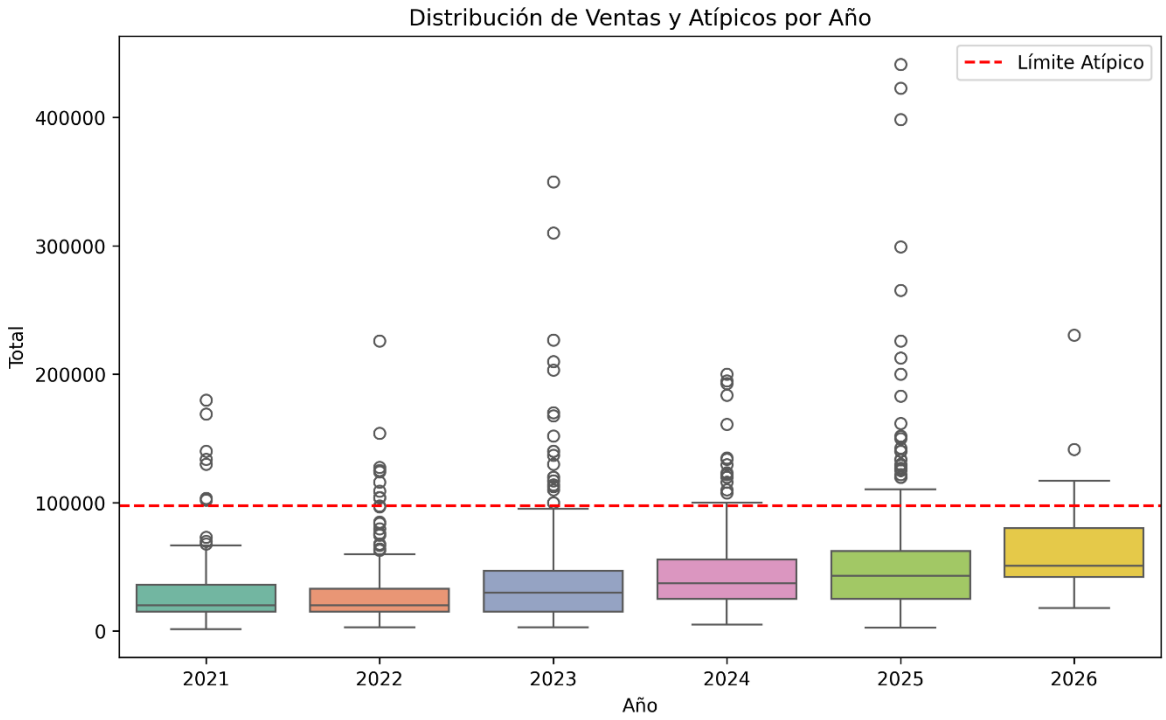
## 3.3. Outliers

El análisis con el método IQR sobre Total detectó una proporción acotada de outliers:

- Los outliers corresponden principalmente a:
    - Ventas de neumáticos en conjuntos (2 neumáticos o más).
    - Paquetes de mantenciones completas de vehículos.
    - Órdenes grandes facturadas a empresas.

Estos outliers no son errores, sino ventas de alto valor coherentes con el negocio. Por eso, se mantuvieron en el dataset, pero se marcaron para analizarlos por separado cuando corresponde.





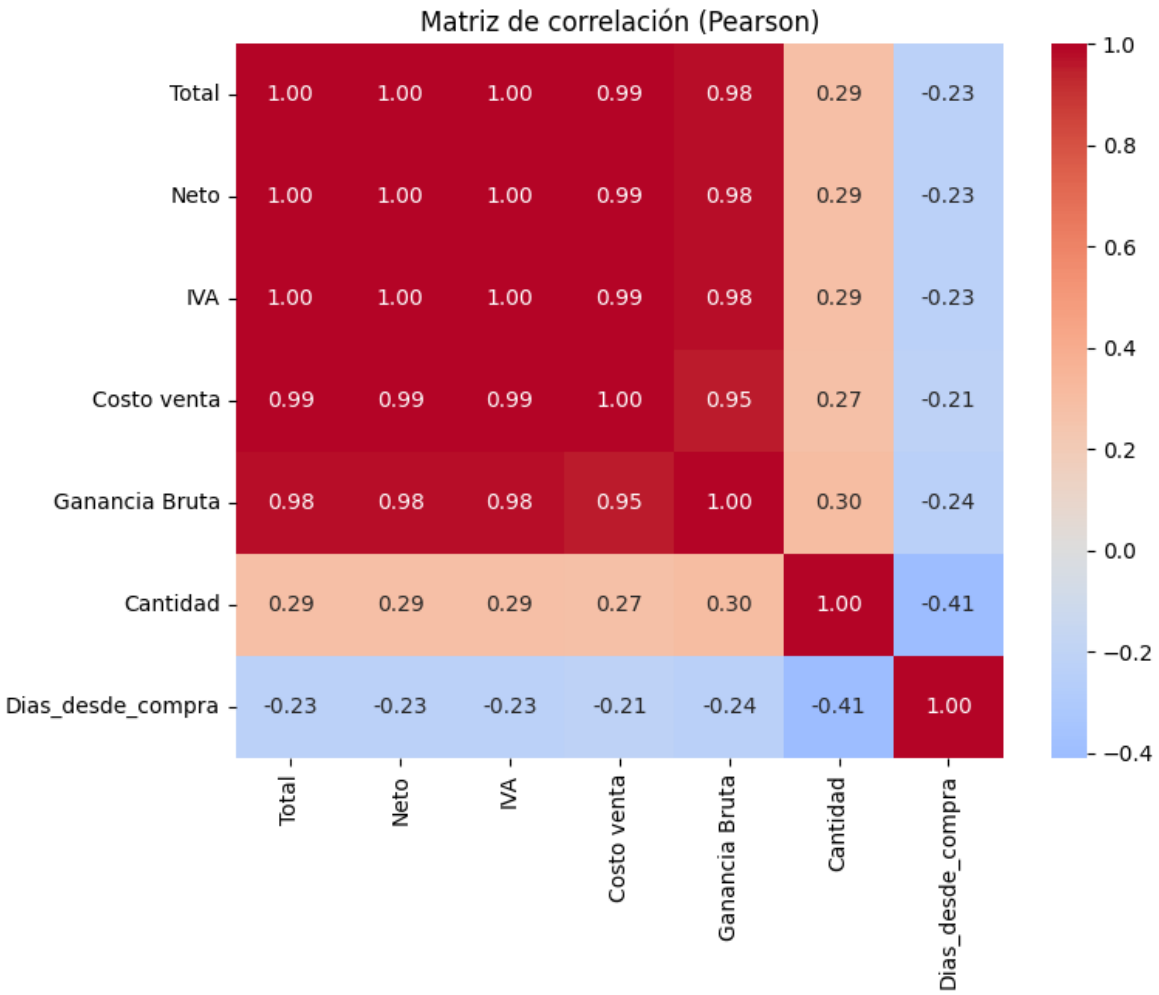

# 4. Correlaciones y relaciones entre variables

Se construyó una matriz de correlación de Pearson entre Neto, IVA, Total, Costo venta, Ganancia Bruta, Cantidad y Dias_desde_compra.

Principales hallazgos:

- **Total, Neto e IVA** presentan correlaciones prácticamente perfectas, lo esperable dado que IVA se calcula a partir del Neto y Total.
- **Total y Costo venta** también muestran correlación muy alta (>0,99), reflejando que a mayores ventas, mayores costos.
- **Total y Ganancia Bruta** presentan una correlación fuerte (~0,98), coherente con un margen porcentual relativamente estable.
- **Ganancia Bruta y Costo venta** muestran correlación alta, dado que ambas dependen de Total.
- **Cantidad** tiene correlación positiva moderada con Total y Ganancia Bruta: las ventas con más ítems tienden a montos mayores, aunque con bastante dispersión.
- **Dias_desde_compra** no presenta una correlación fuerte con montos, lo que indica que no hay una relación lineal clara entre antigüedad de la venta y su tamaño.

Estas correlaciones explican bien la estructura económica de las ventas, pero también implican multicolinealidad cuando se usan todas estas variables juntas en modelos de regresión.




# 5. Modelos de regresión

## 5.1. Regresión lineal simple: Total- Cantidad

Se ajustó un modelo de regresión lineal simple tomando **Total** como variable dependiente y **Cantidad** como predictor (número de ítems aproximado en la descripción del producto):

- El modelo obtuvo un $\mathbf{R^2 \approx 0{,}082}$, lo que indica que la Cantidad explica alrededor del 8,2% de la variabilidad observada en el Total.
- Ambos coeficientes son estadísticamente significativos (p-valor < 0,001 para el intercepto y para Cantidad), con:

    - Intercepto ≈ 22.820: valor esperado de Total cuando Cantidad = 0 (interpretado solo como parámetro del modelo, no como caso real de negocio).
    - Coeficiente de **Cantidad ≈ 10.640**: en promedio, **cada ítem adicional en la venta se asocia a un incremento de alrededor de 10.600 en el Total**, manteniendo el resto constante.
- El bajo $R^2$, junto con un estadístico F significativo, indica que la relación es real pero débil: el monto total de la venta **aumenta con la cantidad de productos**, pero hay muchos otros factores (tipo de producto, tipo de cliente, canal, etc.) que influyen en el valor final.

En términos prácticos, este modelo muestra que las ventas con más ítems tienden a ser más altas, pero la cantidad por sí sola no es suficiente para predecir con precisión el total de la venta.

## 5.2. Regresión lineal simple: Total ~ Costo venta

Se ajustó un segundo modelo de regresión lineal simple con **Total** como variable dependiente y **Costo venta** como predictor:

- El modelo presenta un **$R^2 \approx 0{,}985$**, lo que significa que el Costo de venta explica aproximadamente el 98,5% de la variabilidad en el Total.
- Ambos coeficientes son altamente significativos (p-valor < 0,001):
    - Intercepto ≈ 561,5: valor base del Total cuando el Costo venta es cero (de nuevo, principalmente un parámetro del modelo).
    - Coeficiente de **Costo venta ≈ 1,51**: en promedio, por cada unidad adicional de costo, el Total aumenta en torno a 1,51 unidades.
- Este resultado es coherente con la lógica del negocio y la forma en que se construyeron los datos: a mayores costos, mayores precios de venta, con un margen relativamente estable.

A efectos analíticos, este modelo confirma una **relación casi lineal entre Costo y Total**, consistente con los precios basados en un margen a partir del costo. Sin embargo, la correlación tan alta también implica que, en modelos múltiples, usar simultáneamente Total, Neto, IVA y Costo generará multicolinealidad, por lo que sus coeficientes deben interpretarse con cautela.

## 5.3. Regresión lineal múltiple

También se probó un modelo múltiple con Total como variable dependiente y predictores como Cantidad y Costo venta:

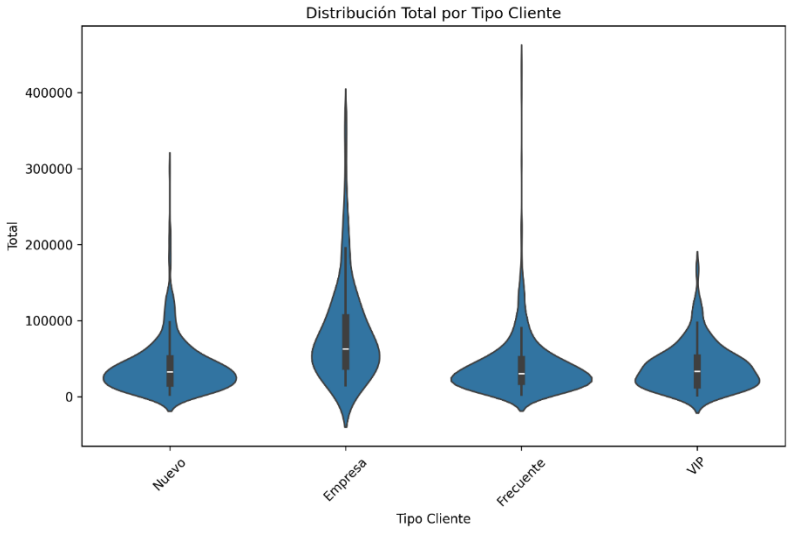
- **$R^2$ muy alto** (cerca de 0,99) por incluir Costo venta.
- **Cantidad** tendrá coeficiente positivo pero bajo impacto ($R^2$ simple era solo 0,082).
- **Costo venta** dominará la explicación (coeficiente ≈1,5).

El modelo múltiple confirma que **Costo venta es el predictor dominante**, mientras que **Cantidad aporta información adicional pero marginal**. Esto es coherente con el negocio: el costo explica casi todo el precio final, y la cantidad solo incrementa el Total de forma secundaria.

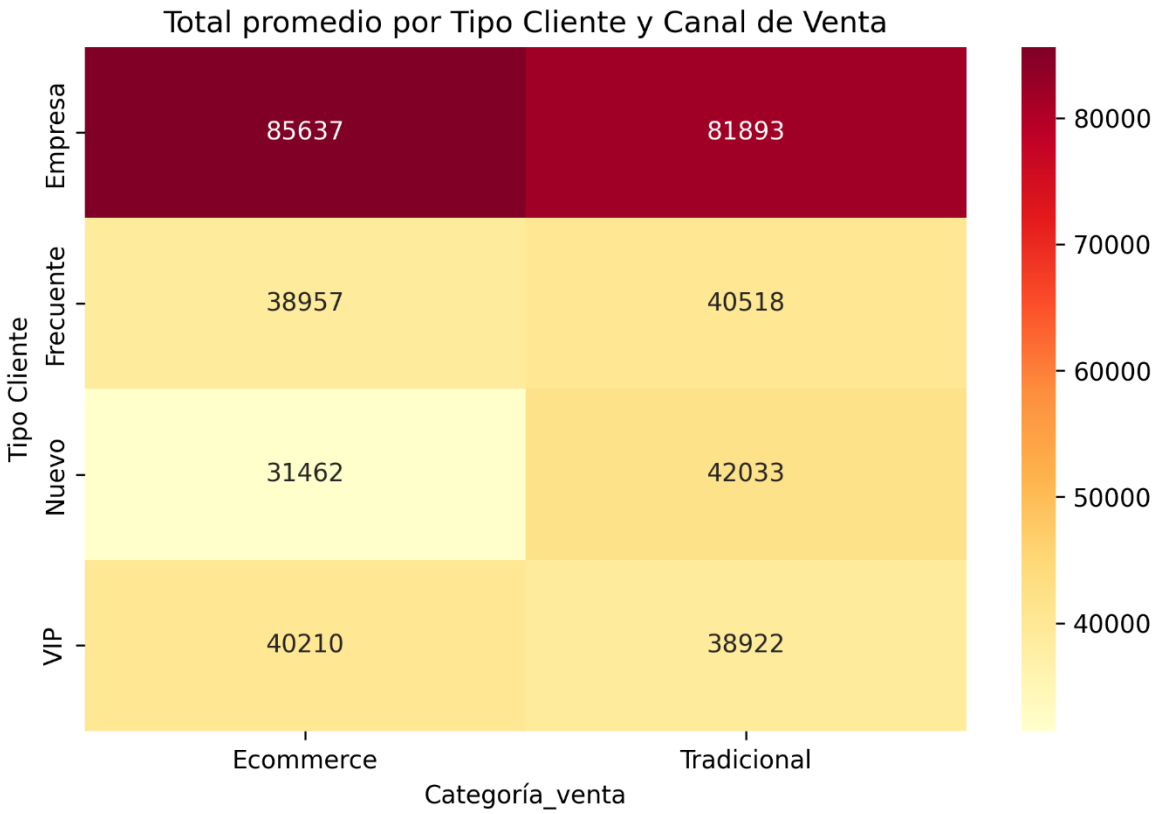
# 6. Análisis visual con Seaborn y Matplotlib

Se construyeron diversas visualizaciones para complementar el análisis numérico:

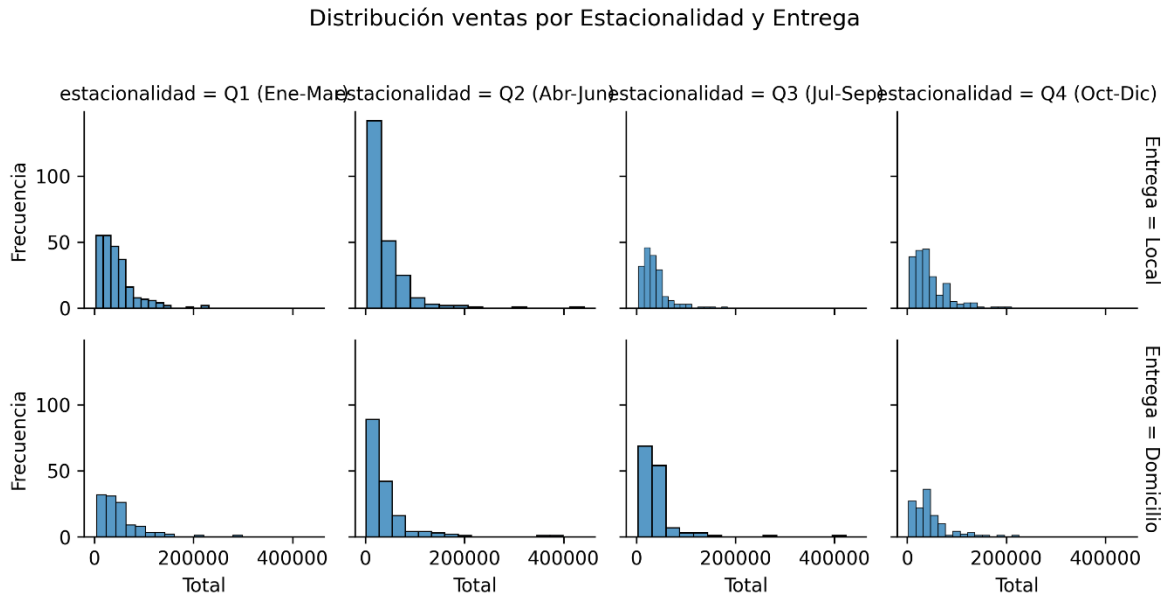
- **Histogramas y boxplots** de Total, Ganancia Bruta, Costo venta: ayudan a ver la concentración y los outliers.
- **Violinplots y boxplots por categoría:**
    - Total por Tipo Cliente: las empresas (Factura) concentran operaciones de mayor valor, mientras que clientes Nuevos/Frecuentes/VIP se ubican en rangos menores pero más numerosos.



- **Heatmaps:**
    - Total promedio por Tipo Cliente y Categoría_venta muestran que:
        - Las ventas B2B (Empresa) tienden a tickets medios más altos.
        - El canal Tradicional concentra el volumen, pero Ecommerce aporta tickets mayores en ciertos segmentos.




- **FacetGrid por estacionalidad y Entrega:**
    - Se observan patrones de mayor distribución de ventas en determinados trimestres, con ventas a Domicilio algo más frecuentes en ciertos períodos, aunque sin estacionalidad extremadamente marcada.




- **Evolución temporal con Matplotlib:**
    - La serie de ventas mensuales muestra una tendencia creciente en el tiempo, con picos asociados a ventas grandes a empresas y paquetes de mantención.

# 7. Hallazgos fundamentales

A partir del análisis realizado, se destacan los siguientes hallazgos clave:

1. **La estructura de ingresos tiene 2 componentes:** una de ventas de bajos montos y muy frecuentes y otra de montos altos, pero de frecuencia más baja.
2. **Importancia del segmento empresas (Factura):** las empresas concentran los mayores montos por transacción. Estas ventas, aunque menos frecuentes son una parte importante del ingreso total de la empresa, por lo que es recomendable concentrar esfuerzos de marketing en este segmento.
3. **Clientes retail segmentados por recurrencia:** clientes frecuentes sostienen gran parte del volumen de ventas, los clientes VIP en tanto en promedio tienden a tickets más altos.
4. **Márgenes y rentabilidad razonables:** los márgenes brutos se mantienen entre el 30% y 40%, que es estable, pero hay espacio para optimizar según segmentos específicos.

5. **Outliers que corresponden a oportunidades, no a errores:** No se identificaron outliers negativos. En los outliers más altos se representan ventas relevantes y no errores o anomalías por corregir.

# 8. Conclusiones y recomendaciones

En conclusión, el análisis exploratorio del registro de ventas de la empresa muestra un negocio que tiene una operación con márgenes de ganancia razonables para el tipo de productos, con ventas muy marcadas entre ventas de pequeño monto, pero mucha frecuencia y ventas de poca frecuencia, pero monto alto que generalmente corresponde a compras de empresas. Esto representa una oportunidad para el negocio, que debería ser abordada por su área comercial.

A partir de estos hallazgos, se proponen las siguientes recomendaciones:

1. **Profundizar en el segmento empresas (Factura):**
    - Identificar explícitamente clientes empresa de mayor contribución (top N por facturación anual).
    - Diseñar ofertas o condiciones especiales que consoliden estas relaciones y fomenten la recurrencia.
    - Buscar la captación de nuevos clientes empresa, esto fundamentado en el alto monto de las ventas asociadas a este tipo de cliente.
2. **Fortalecer la fidelización de clientes frecuentes y VIP:**
    - Implementar comunicaciones o beneficios diferenciados para estos segmentos.
    - Monitorear su frecuencia de compra y ticket medio en el tiempo, para anticipar caídas de actividad.
3. **Monitorear outliers como oportunidades estratégicas:**
    - No eliminar estas ventas del análisis; por el contrario, analizarlas como casos de éxito: tipos de productos involucrados en estas ventas y tipo de cliente que realizó la compra.
4. **Próximos pasos analíticos:**
    - Integrar información adicional de clientes (ubicación, rubro, frecuencia histórica real) para modelos más ricos.

- Desarrollar modelos de predicción de ventas por cliente/segmento y análisis de canasta de productos.
- Automatizar reportes mensuales con los principales indicadores: ventas totales, margen, top productos, top clientes y análisis de outliers.

Mediante este análisis es posible comprender con mayor profundidad el comportamiento de ventas de una empresa y como las distintas variables trabajan para lograr los buenos resultados. Si bien algunos resultados se simularon para mejorar el análisis se pudieron agregar más columnas numéricas para ver como se correlacionaban con las otras variables.