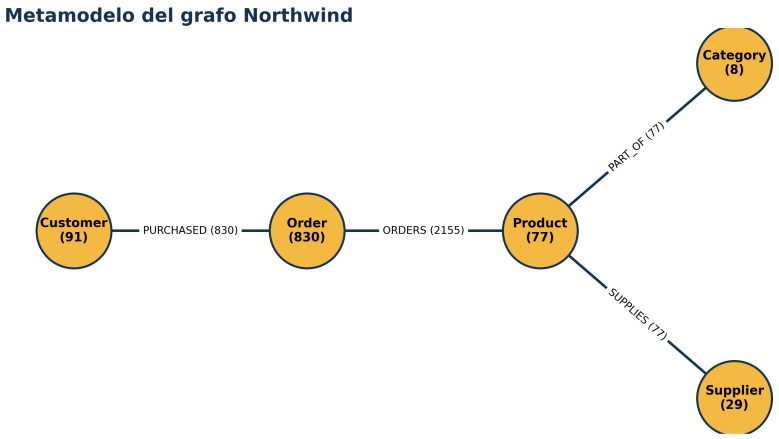
Reporte Ejecutivo

Analisis del Catalogo de Productos y Comportamiento de Clientes

Caso Northwind Retail



Nombre	Esteban Garzon
Asignatura / Actividad	Actividad 4 - Exploracion Analitica de la Base Northwind Retail
Rol asumido	Cientifico de datos especializado en analitica de grafos y aprendizaje automatico
Entregable	Informe ejecutivo en LaTeX con evidencia visual y scripts reutilizables
Archivos adjuntos	`consultas_northwind.cypher` y `analyze_northwind.py`

13 de marzo de 2026

1. Resumen Ejecutivo

Se reconstruyo el grafo Northwind desde el export oficial `northwind-50.cypher` y se analizaron **1.035 nodos** y **3.139 relaciones**. La estructura contiene **91 clientes**, **830 ordenes**, **77 productos**, **29 proveedores** y **8 categorias**. El metamodelo confirma que el negocio se organiza alrededor de cuatro flujos principales: cliente→orden, orden→producto, producto→categoria y proveedor→producto.

Los principales hallazgos fueron:

- La concentracion geografica del archivo `reporte_compras.csv` se ubica en **USA, Germany y Austria**; por ciudad dominan **Boise, Graz y Cunewalde**.
- En terminos monetarios, las categorias con mayor venta son **Beverages** (267.868,18) y **Dairy Products** (234.507,28); por pais lideran **USA** y **Germany**.
- El producto mas influyente por *PageRank* es **Côte de Blaye** (0,0430) y el cliente mas central es **Jose Pavarotti** (0,0367).
- La red de co-compra es muy densa: **Roland Mendel** alcanza 88 conexiones y la comunidad Louvain mas grande agrupa 24 clientes.
- Existe riesgo de reposicion en productos de amplio alcance como **Gorgonzola Telino**, **Gnocchi di nonna Alice**, **Queso Cabrales** y **Outback Lager**.

Nota metodologica. En el entorno entregado no existia una instancia Neo4j Browser / AuraDB activa. Para mantener trazabilidad, las consultas Cypher se documentan y se dejaron listas para ejecutar en `consultas_northwind.cypher`, mientras que los resultados mostrados se reprodujeron sobre el mismo export oficial del grafo.

2. Diseno del Grafo

2.1. Metamodelo

La Figura 1 resume el modelo conceptual del grafo. Esta representacion es adecuada para retail porque conecta comportamiento de compra, portafolio y abastecimiento sin necesidad de tablas puente complejas.

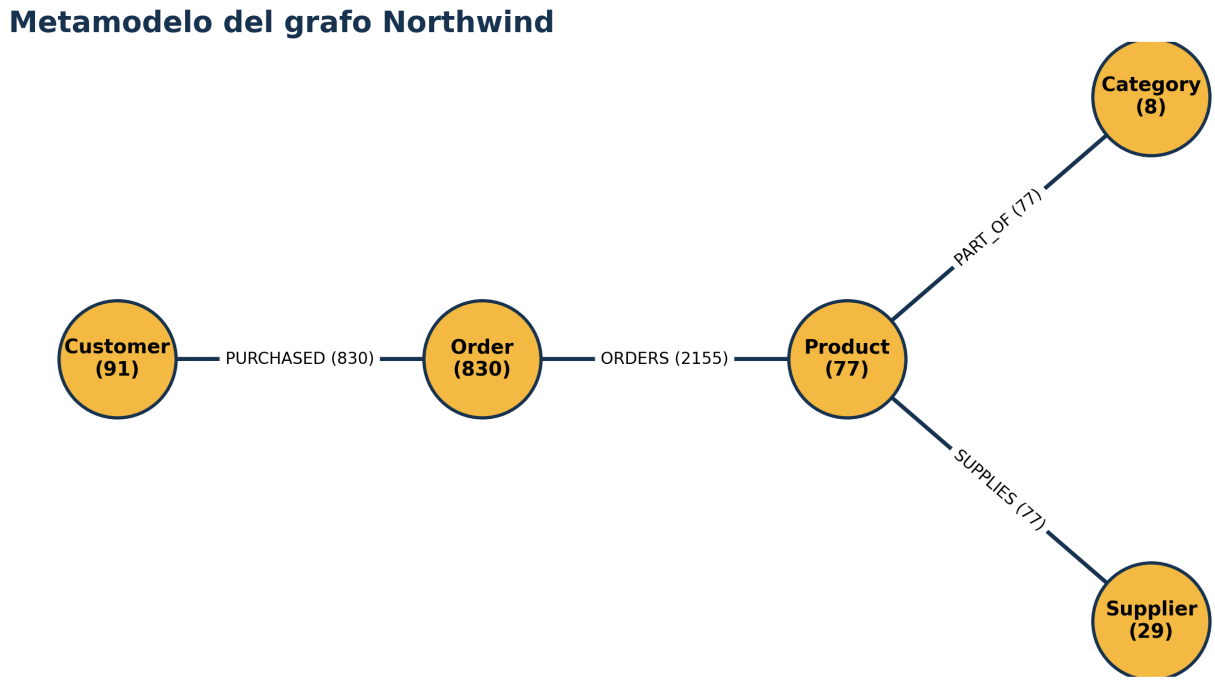


Figura 1: Metamodelo del grafo Northwind.

Tipos de nodos

tipo_nodo	conteo
Category	8
Customer	91
Order	830
Product	77
Supplier	29

(a) Tipos de nodos

Tipos de relaciones

tipo_relacion	conteo
ORDERS	2155
PART_OF	77
PURCHASED	830
SUPPLIES	77

(b) Tipos de relaciones

Figura 2: Evidencia estructural del grafo.

2.2. Tipos de nodos y relaciones

Customer: representa el cliente final; incluye identificador, nombre de contacto, ciudad, pais y datos de empresa. **Order**: concentra el evento transaccional con fecha, destino de envio y flete. **Product**: describe precio, inventario, estado de descontinuacion y proveedor/categoria asociada. **Category**: agrupa productos por familia comercial. **Supplier**: identifica al proveedor y permite conectar demanda con abastecimiento.

Las relaciones tienen semantica clara: **PURCHASED** enlaza cliente y orden, **ORDERS** enlaza orden y producto con atributos transaccionales (*unitPrice*, *quantity*, *discount*), **PART_OF** asocia productos a categorias y **SUPPLIES** modela el abastecimiento desde proveedor.

2.3. Consultas base en Cypher

Listing 1: Conteo de nodos y relaciones del metamodelo

```
MATCH (n)
RETURN labels(n)[0] AS tipo_nodo, count(*) AS conteo
ORDER BY conteo DESC;

MATCH ()-[r]->()
RETURN type(r) AS tipo_relacion, count(*) AS conteo
ORDER BY conteo DESC;
```

3. Consultas Analiticas

3.1. Top 5 mensual de clientes

Se construyeron dos variantes: una por **volumen** para alinear la salida con `reporte_compras.csv`, y otra por **valor monetario** para la toma de decisiones gerenciales. La consulta siguiente devuelve el ranking mensual por ventas:

Listing 2: Top 5 mensual por valor monetario

```
MATCH (c:Customer)-[:PURCHASED]->(o:Order)-[od:ORDERS]->(:Product)
WITH date(datetime(o.orderDate)) AS fecha,
     c.contactName AS cliente, c.country AS pais, c.city AS ciudad,
     sum(toFloat(od.unitPrice) * toInteger(od.quantity) *
         (1 - toFloat(coalesce(od.discount, 0)))) AS venta_total
ORDER BY fecha.year, fecha.month, venta_total DESC
WITH fecha.year AS anio, fecha.month AS mes,
     collect({cliente: cliente, pais: pais, ciudad: ciudad,
              venta_total: round(venta_total * 100) / 100.0}) AS filas
UNWIND range(0, CASE WHEN size(filas) < 5 THEN size(filas) - 1 ELSE 4 END) AS idx
RETURN anio, mes, idx + 1 AS ranking, filas[idx] AS fila;
```

Top 5 de clientes por mes

year	month	customer_name	country	city	total_comprado
1996	7	Pascale Cartrain	Belgium	Charleroi	3597.9
1996	7	Peter Franken	Germany	München	3536.6
1996	7	Roland Mendel	Austria	Graz	3488.68
1996	7	Mario Pontes	Brazil	Rio de Janeiro	2997.4
1996	7	Michael Holz	Switzerland	Genève	2490.5
1996	8	Horst Kloss	Germany	Cunewalde	6796.64
1996	8	Paula Wilson	USA	Albuquerque	3343.6
1996	8	Pedro Afonso	Brazil	Sao Paulo	2169
1996	8	Christina Berglund	Sweden	Luleå	2102
1996	8	Renate Messner	Germany	Frankfurt a.M.	1521.375
1996	9	Paula Wilson	USA	Albuquerque	4929.3
1996	9	Patricia McKenna	Ireland	Cork	4407

Figura 3: Captura del resultado tabular del top 5 mensual por valor monetario.

Interpretacion: los clientes **Horst Kloss**, **Roland Mendel** y **Jose Pavarotti** dominan el valor facturado a lo largo del periodo. Esto sugiere una alta dependencia de pocos compradores premium, util para estrategias de fidelizacion y priorizacion comercial.

Hallazgo adicional: al contrastar el resultado con `reporte_compras.csv`, se detecto que ese CSV representa principalmente **cantidades compradas** y no importe monetario. Por ejemplo, en julio de 1996, Roland Mendel registra **305** en el CSV, mientras su valor monetario asciende a **3.488,68**. Esta distincion es clave para evitar decisiones basadas solo en volumen.

4. Visualizacion Gerencial

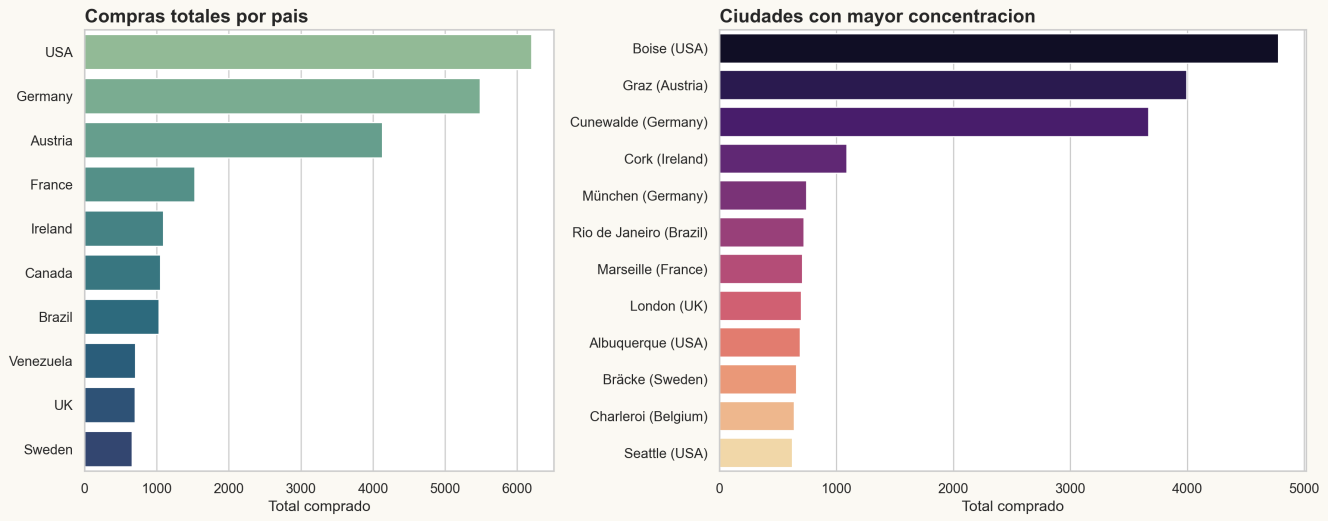


Figura 4: Distribucion de compras totales por pais y ciudad a partir de `reporte_compras.csv`.

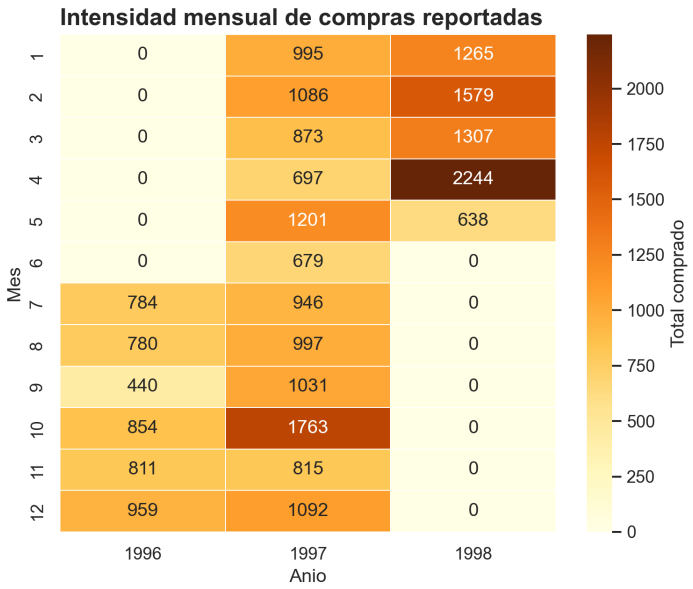


Figura 5: Intensidad mensual de compras reportadas.

La evidencia geografica muestra una concentracion fuerte en **USA** (6.202), **Germany** (5.486) y **Austria** (4.130). A nivel ciudad, **Boise**, **Graz** y **Cunewalde** explican una parte desproporcionada del volumen. Desde el punto de vista ejecutivo, esto habilita dos decisiones: focalizar campañas en polos de alta respuesta y revisar si la dependencia geografica implica riesgo comercial.

5. Busqueda y Algoritmos de Grafos

5.1. Shortest path

La consulta pedida identifica rutas minimas entre pares de clientes, considerando hasta cinco saltos.

Listing 3: Consulta de shortest path entregada en el enunciado

```
MATCH (a:Customer)-[*..5]-(b:Customer)
WHERE elementId(a) < elementId(b)
MATCH path = allShortestPaths((a)-[*..5]-(b))
RETURN a.contactName AS clienteA, b.contactName AS clienteB,
       length(path) AS saltos, path
LIMIT 10;
```

Muestra de shortest paths

clienteA	clienteB	saltos	ruta
Maria Anders	Ana Trujillo	2	Maria Anders -> Antonio Moreno -> Ana Trujillo
Maria Anders	Antonio Moreno	1	Maria Anders -> Antonio Moreno
Maria Anders	Thomas Hardy	1	Maria Anders -> Thomas Hardy
Maria Anders	Christina Berglund	1	Maria Anders -> Christina Berglund
Maria Anders	Hanna Moos	1	Maria Anders -> Hanna Moos
Maria Anders	Frédérique Citeaux	1	Maria Anders -> Frédérique Citeaux
Maria Anders	Martín Sommer	2	Maria Anders -> Antonio Moreno -> Martín Sommer
Maria Anders	Laurence Lebihan	1	Maria Anders -> Laurence Lebihan
Maria Anders	Elizabeth Lincoln	1	Maria Anders -> Elizabeth Lincoln
Maria Anders	Victoria Ashworth	1	Maria Anders -> Victoria Ashworth

Figura 6: Muestra de rutas mas cortas entre clientes.

Interpretacion: la mayoria de pares mostrados se conectan en uno o dos saltos, lo que evidencia un grafo comercial compacto. En retail, rutas cortas facilitan encontrar clientes analogos para *cross-sell*, recomendacion

de productos o expansion de segmentos.

5.2. Construccion de la red de co-compra y centralidad por grado

Primero se materializa una relacion `CO_PURCHASED` cuando dos clientes compran al menos un producto comun; luego se cuentan sus conexiones.

Listing 4: Construccion de CO_PURCHASED y consulta de conexiones

```
MATCH (c1:Customer)-[:PURCHASED]->(:Order)-[:ORDERS]->(p:Product)
      <-[:ORDERS]-(:Order)<-[:PURCHASED]-(c2:Customer)
WHERE elementId(c1) < elementId(c2)
MERGE (c1)-[:CO_PURCHASED]->(c2);

MATCH (c:Customer)-[:CO_PURCHASED]-()
RETURN c.contactName AS cliente, count(*) AS conexiones
ORDER BY conexiones DESC
LIMIT 10;
```

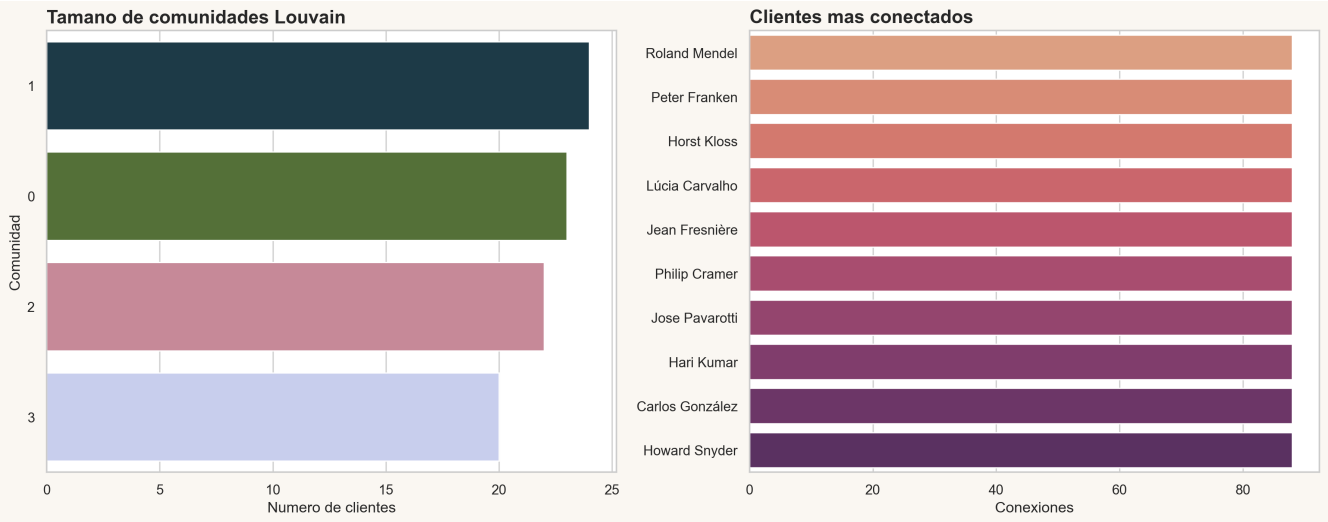


Figura 7: Tamano de comunidades Louvain y clientes mas conectados por co-compra.

Clientes con mas conexiones CO_PURCHASED

customer_node	cliente	pais	ciudad	conexiones
C::ERNSH	Roland Mendel	Austria	Graz	88
C::FRANK	Peter Franken	Germany	München	88
C::QUICK	Horst Kloss	Germany	Cunewalde	88
C::QUEEN	Lúcia Carvalho	Brazil	Sao Paulo	88
C::MEREP	Jean Fresnière	Canada	Montréal	88
C::KOENE	Philip Cramer	Germany	Brandenburg	88
C::SAVEA	Jose Pavarotti	USA	Boise	88
C::SEVES	Hari Kumar	UK	London	88
C::LILAS	Carlos González	Venezuela	Barquisimeto	88
C::GREAL	Howard Snyder	USA	Eugene	88
C::TORTU	Miguel Angel Paolino	Mexico	México D.F.	87
C::LINOD	Felipe Izquierdo	Venezuela	I. de Margarita	87

Figura 8: Captura del ranking de clientes mas conectados.

Interpretacion: **Roland Mendel**, **Peter Franken**, **Horst Kloss** y otros clientes premium alcanzan **88 conexiones**. Esto equivale a una centralidad por grado muy alta y los convierte en candidatos naturales para pruebas piloto, programas VIP o lanzamiento de nuevos productos.

5.3. Productos con mayor alcance, PageRank y Louvain

La consulta de alcance de producto mide cuantos clientes distintos tocan cada producto. Sobre la red reconstruida tambien se calculo *PageRank* y deteccion de comunidades tipo Louvain.

Listing 5: Productos con mayor numero de clientes

```
MATCH (c:Customer)-[:PURCHASED]->(:Order)-[:ORDERS]->(p:Product)
WITH p, collect(DISTINCT c.contactName) AS clientes
RETURN p.productName AS producto, clientes, size(clientes) AS num_clientes
ORDER BY num_clientes DESC
LIMIT 10;
```

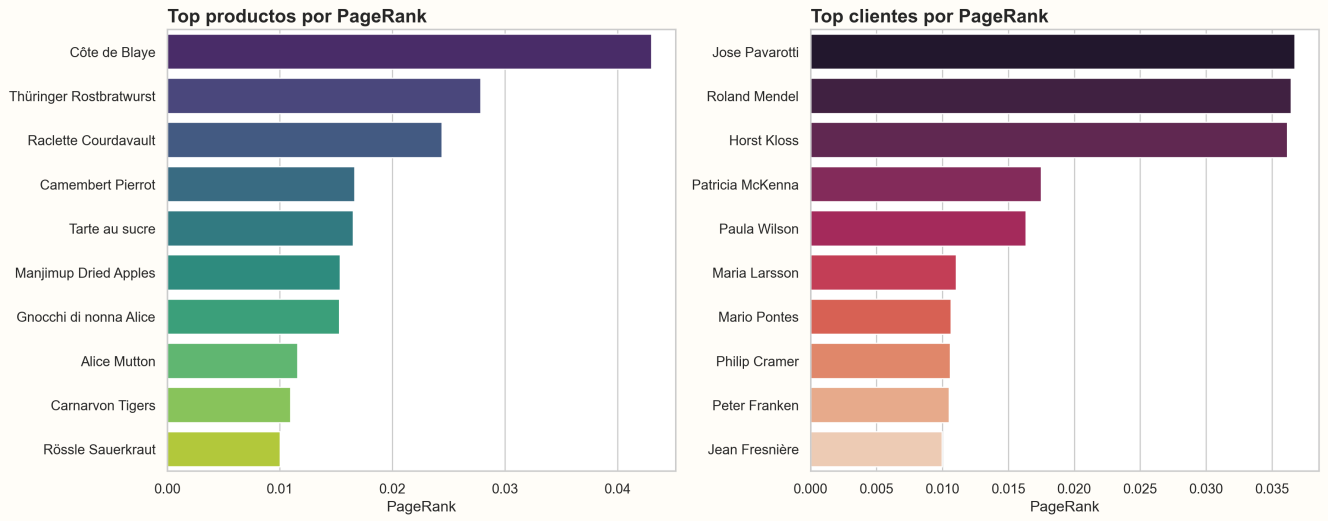


Figura 9: Top productos y clientes por PageRank.

(a) Productos con mayor alcance

(b) Top nodos por PageRank

Figura 10: Resultados destacados de alcance y centralidad.

Interpretacion:

- **Raclette Courdavault** es el producto con mayor cobertura (**43 clientes**), seguido de **Camembert Pierrot** y **Gnocchi di nonna Alice**. Son productos ideales para *bundles* o promociones ancla.
- El *PageRank* mas alto en productos pertenece a **Côte de Blaye**; su relevancia no depende solo del numero de compradores, sino de estar conectado a clientes influyentes.
- Louvain encontro **4 comunidades** relativamente balanceadas (24, 23, 22 y 20 clientes), lo que sugiere segmentos por afinidad de canasta mas que por geografia pura.

Tipo de algoritmo y utilidad practica. *Shortest path* es un algoritmo de caminos minimos; *PageRank* es una medida de importancia topologica; *Louvain* es un algoritmo heuristico de deteccion de comunidades. En conjunto permiten recomendar productos, identificar clientes semilla y descubrir microsegmentos.

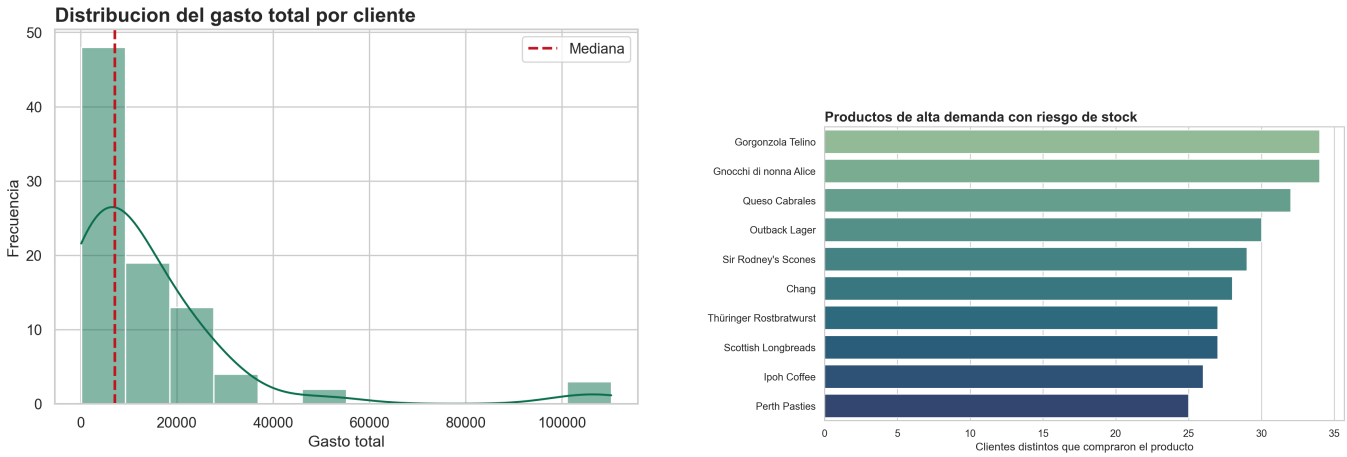
6. Aprendizaje Automatico

6.1. Pregunta de negocio

Clasificacion: "*Que clientes tienen mayor probabilidad de convertirse en clientes de alto valor en el siguiente periodo?*"
Regresion: "*Cuanto gastara cada cliente en su siguiente ventana mensual?*"

6.2. Enfoque propuesto

Se construyo un dataset por cliente con variables transaccionales y de red: gasto total, ticket promedio, numero de ordenes, numero de productos y categorias, descuento promedio, *PageRank*, grado en la red `CO_PURCHASED` y comunidad Louvain. La Figura 11a muestra la distribucion del objetivo.



(a) Variable objetivo de gasto

(b) Soporte a decisiones de inventario

Figura 11: Variables de negocio para modelado y accion.

Preparacion de datos recomendada:

- Depurar nulos y normalizar textos de pais/ciudad.
- Convertir fechas, precios, cantidades y descuentos a formato numerico consistente.
- Codificar comunidades y paises con *target encoding* o *one-hot encoding*.
- Estandarizar variables continuas si se usa regresion lineal o SVM.
- Separar entrenamiento/validacion por tiempo para evitar fuga de informacion.

Algoritmos sugeridos:

- **Clasificacion**: *XGBoost* o *Random Forest*, porque capturan relaciones no lineales y mezclan bien variables transaccionales con metricas de red.
- **Regresion**: *Gradient Boosting Regressor* para estimar gasto futuro con robustez ante distribuciones sesgadas.

Estas variables de grafo suelen aportar valor incremental porque condensan afinidad, influencia y posicion estructural del cliente, factores que no aparecen directamente en un modelo relacional tradicional.

7. Conclusiones y Recomendaciones

El grafo Northwind demuestra que la relacion cliente-producto-proveedor es suficientemente rica para soportar decisiones comerciales, de segmentacion e inventario. El negocio no solo depende de cuantos clientes compran, sino de **quienes** estan conectados, **que** productos articulan la red y **donde** se concentra la demanda.

Recomendaciones gerenciales:

- Priorizar programas de fidelizacion y venta consultiva sobre **Horst Kloss**, **Roland Mendel** y **Jose Pavarotti**, porque concentran valor y centralidad.
- Diseñar promociones ancla con **Raclette Courdavault**, **Camembert Pierrot** y **Gnocchi di nonna Alice**; son productos de alta difusion en la red.

- Reforzar abastecimiento de **Gorgonzola Telino**, **Gnocchi di nonna Alice**, **Queso Cabrales** y **Outback Lager**; combinan amplio alcance con inventario tensionado.
- Revisar sustitutos para productos descontinuados pero aun estructuralmente relevantes, como **Thüringer Rostbratwurst** y **Alice Mutton**.
- Ejecutar campañas por afinidad de canasta usando comunidades Louvain, no solo por geografia, ya que los segmentos detectados son balanceados y transversales.

Scripts utilizados. El detalle tecnico queda disponible en `consultas_northwind.cypher` (consultas Cypher listas para Neo4j) y `analyze_northwind.py` (reconstruccion del grafo, metricas, tablas y visualizaciones).